IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received October 17, 2019, accepted October 28, 2019, date of publication November 4, 2019, date of current version November 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2951164

Data-Driven Logical Topology Inference for Managing Safety and Re-Identification of Patients Through Multi-Cameras IoT

KEYANG CHENG[1], MUHAMMAD SADDAM KHOKHAR[1], QING LIU[1], RABIA TAHIR[1], AND MAOZHEN LI[2], (Member, IEEE)

[1]School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
[2]Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, U.K.

Corresponding author: Keyang Cheng (kycheng@ujs.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61972183 and Grant 61602215, and in part by the Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by the Big Data (PSRPC).

ABSTRACT As Internet of Things (IoT) develops, IoT technologies are starting to integrate intelligent cameras for managing safety within mental health hospital wards and relevant spaces, seeking out specified individuals from these surveillance videos filmed by the various cameras. Because monitoring is one of the important application of IoT based on distributed video cameras. In order to fine-grained re-identification of patients and their activities against the very low resolution, occlusions and pose, viewpoint and illumination changes, we propose a novel data-driven model to infer multi-cameras logical topology and re-identify patients captured by different cameras. In our model, we employ a Time-Delayed Mutual Information (TDMI) model in order to address multi-cameras logical topology inference. Additionally, we use a well-trained Deep Convolutional Neural Network (DCNN) to extract characteristics. Moreover, we employ a name-ability model to discover deep attributes and a classifier based on a structural output of attributes is designed to tackle the re-identification of patients, especially who possess psychiatric behaviour. In order to improve the present model's performance, we resort to the parallelized implementations. Experimental results show that our model possesses the best performance as compared to state-of-the-art model,especially, when the semantic restrictions are imposed onto the production of patients' specific attributes with structural output. Further, the deep learning model is used to produce characteristics when there is no supervision on the learning model of attributes.

INDEX TERMS Canonical correlation analysis, time delayed mutual information (TDMI), deep convolutional neural network (DCNN), multi-camera topology inference.

I. INTRODUCTION

From last two decades, videos surveillance systems have been used for managing safety within healthcare industry especially in mental health hospital, asylum, seclusion room, and wards for controlling the suspicious activities of psychotic patients. Such as spotting suspicious behavior, violence, hyperactivity, and safety issues. Further, these issues lead to treatment-related harms caused by medical care of patient and staff. Additionally, those adverse events also lead to harm caused by errors, medical treatment errors, potential for harm. further, multi-cameras monitoring surveillance system can sort the problems related to managing medical care protocols [1]–[3]. Health monitoring research leads to significant focus to improve and achieve good healthcare environment. The analytical evaluation of mentioned issues and analyzing their causes can help in healthcare planning and design advance smart health care system too. Unfortunately, most hospitals and health systems have rudimentary methods to identify this issues or activities. These rudimentary methods mostly depend on the use of voluntary help, chart-reports results, and scanning of manually observed behavior of patient. These methods face challenges and limitations those lead them to accidents. Further, these methods required human resources

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Cheng.

as well. Detecting these events is especially challenging because of the unpredictable nature of patients. Along with powerful re-identification features, this paper proposes a model to improve patient safety in hospitals through the usage of Multi-Cameras IoT. With the advent of Cloud technology, the notion of connecting any and every device to the Internet with an on and off switch has become a reality. In the last few years, the Internet of Things has taken off and grown substantially across the world. Internet is no longer confined to computers and mobile devices; it is now available to nearly every device that has an IP address from microwaves, refrigerators to wearable devices and headphones. As the Internet of Things (IoT) is being developed, IoT technologies are beginning to integrate intelligent cameras. Distributed video cameras have played important roles in various IoT applications, especially in the analysis of inter-camera entity association, or re-identification of person, which serves as one of the most significant applications of IoT based on distributed video cameras. In mental health hospitals, new surveillance technology has a significant impact of monitoring not only psychotic patient but monitoring staff as well. The proposed model brings more intelligent observation of human behavior in monitoring system as well. It has a potential for influencing mental health practice with in a closed environment in many ways. Required by managing safety within healthcare industry, patient re-identification is a challengingtask to identify patients in various scenes obtained by the non-overlapping cameras in different spaces. Because these monitoring video surveillance applications performed through both time and space, the patients vanishing in a scene is supposed to make a distinction with an enormous number of probable objects and match with only one or multiple scenes in diverse situations as well as at a different point in time. It is likely that every view could be capture by one diverse perspective, with various conditions of lighting in both dynamic and static states, and the degrees of occlusion as well as different particular variables of view. In face recognition, Ou *et al.* [4] propose a more general approach based on structured sparse representation for face recognition with contiguous occlusion and he also propose [5] a robust nonnegative patch alignment (RNPA) for dimensionality reduction. To deal with the similar task, we obtain the logical topology of distributed cameras first, which helps us to set the order of searching. Some latest works of cameras logical topology inference based on multiple camera activity correlation analysis have been introduced, in which they discovered modeling correlations between activities in a crowded space with uncelebrated and non-overlapping cameras.

Surveillance cameras began to emerge as a security tools after zero tolerance campaign supported by the hospitals staff receiving end of violence from patients in 2006. According to the research of Winstanley and Whittington (2002), that staffs are more vulnerable and likely to be at risk who works within Psychiatric ward than staff of general hospitals. Further, mental health hospitals are increasingly becoming unsafe and sometimes becoming dangerous place not only for staff but for patients as well. Hartocollis (2008) explained the abuse of patients from the hospital staff that can cause death or violent behavior from psychiatric patients.By deploying the surveillance in mental hospitals, several benefits could be achieved. (1) It could be helpful in nursing practice, because staff can watch patients by cameras. (2) Unpredictable violence could be detected that can lead to harm patient himself or other staff as well. Mental health research suggests that patients of schizophrenia or psychotic illness are more likely to have symptoms of unpredictable violent behavior (Swanson et al, 1996; Taylor, 1997; Taylor and Gunn, 1999). It is not possible to respond nursing staff immediatelythat are being attacked in real time.

With the rapid growth of camera networks of the healthcare industry in recent years, more interest in the community of computer vision is increasing in order to formulate the automatic re-identification solutions. There are two strategies focused primarily by those efforts: (1) developing feature representations that are easy to discriminate the identityas well as invariable to both lighting and view angle [6]–[8]; learning approaches in order to make discrimination and optimize the parameters of a re-identification model [9]. Up to now, the automatic re-identification is still an outstanding issue because of the potential challenge that a majority of visual features do not possess sufficient discrimination for the entity association between cross and view particularly with the images of low resolution or insufficient robustness to view the changes of condition.

While performing re-identification of patients, these human experts usually lean on the matched appearance of the functional properties which are in continuous and distinctive interpretation, such as the styles of hair, shoes, and uniforms [10]. These interpretation are opposite to the continual and obscure quantities measured by the re-identification methods based on modern computer vision utilizing visual characteristics such as texture and color [6]–[8]. However, despite of all patient identification methods, there are still some questions that arise in this area.For example, how to obtain attributes automatically, how to finely describe the humans attributes, how to integrate non-semantical feature extraction of low grade with the semantic attribute classification of middle level and adopt the model to tackle the re-identification of patients. These all challenges arise in re-identification of patients that ought to be handled. In this paper, taking inspiration from recent research [11], we first employ cross-canonical correlation analysis to obtain the cameras logical topology, which will help us to set the order of searching. Then, a classifier based on the structural output of attributes is designed to address the re-identification of patients. The major objective of our work is to use a way driven by data to infer the multiple cameras logical topology and re-identify patients with a structural output of fine-grained attributes. The contributions of our model are as follows:

(1) We employ Time Delayed Mutual Information (TDMI) model to obtain the cameras logical topology for patient

re-identification. (2) A novel model is presented for patients' attributes that are comprehensible and differentiable according to the data-driven. (3) The structural output of attributes gives a fine-grained description of attributes, which not only helps in classifying attributes accurately but can satisfy the demands of human as well, who want to know the location information of attributes. (4) We employ parallelized implementations to improve our model, which contains data parallelism, model parallelism and attribute parallelism.

The main motivation of our model is Multi-Cameras IoT based monitoring system for health care industries because a large number of patients who require constant supervision in hospitals wards, seclusion room, and communal areas for controlling the behavior and environment, such as deter vandalism and other criminal acts. The proposed model increases safety for patients and prevents intruders from gaining access to restricted areas. Further, its continuous real-time monitoring undesired movement increases overall security and safety usage of high end centralized surveillance systems is allowing practical issues faced by staff to prevent aggressive patients from harming themselves, other patients or staff, coercive measures and foremost. The monitoring patient is an important part of healthcare same as treatments. Further, it confirms thatstaffs meet health and safety protocols besides itcan be allowed for remote monitoring using smartphones or tablet for identifiable information.

In this paper, the relevant work is investigated in person re-identification research area and experiment is implemented on PETA Dataset because of the similarity of the problem as per our literature investigation. We cannot find private space data such as hospitals, asylum space, because of privacy concern. Then, we show how to use Time Delayed Mutual Information (TDMI) model to obtain the cameras logical topology. Next, we employ a name-ability model to discover properties and utilize Deep Convolutional Neural Network (DCNN) to obtain the characteristics of patients' images. A Distributed Deep Attribute Learning Model (DDAL) based on the structural output of attribute is design to tackle the re-identification of patients. The model combines non-semantic feature extraction of low level by unsupervised deep learning with the semantic attribute classification of middle level by attribute learning which is supervise. Finally, our model's performance with multi-GPUs parallel integration is show in the experiment.

II. RELATED WORK

A. MULTI-CAMERAS LOGICAL TOPOLOGY INFERENCE

For multi-cameras logical topology inference, there have been a few attempts. Cho *et al.* [12] presented a unified framework which solved the multi-cameras topology problems approximately in different scenarios that a unified model followed different algorithms and filters. Jain *et al.* [13] presented a delay model which holds and sorts patterns errors between multiple cameras to address cameras logical topology inference. Wolf and Schlessman [14] model cameras logical topology inference by utilizing the Distributed Smart Cameras (DSCs). The model can deal with the dependency of the first order regardless of the relationships of arbitrary order. Chu *et al.* [15] adopted the tracking of multi-cameras, so that the trajectories in the view of every camera view could be obtained and divided into the global activities (features) by utilizing the topic models lengthened by Principal Component Analysis (PCA) and color histograms and texture histograms. The Co-occurrence relationship among various activities with a settled temporal threshold was being model. Especially, Loy *et al.* [16] who presented a way to understand multi-cameras activity by time-Delayed correlation analysis gives us an enlightening method to deal with multi-cameras logical topology inference.

B. PERSON RE-IDENTIFICATION

The multi-cameras monitoring system is widely used in various sizes of public and private places. The safety of these public and private places depends heavily on the installation of the monitoring system in various regions. The suspicious person's whereabouts can be tracked and the truth of what happened can also be understood by monitoring a large number of cameras of the system. If we only rely on manual processing of these massive surveillance videos, the efficiency will be extremely low, and the labor costs are quite expensive. R Prates and W Schwartz showed in their work [17] that how person re-identification is an effective solution to these problems. Person re-identification is a process of detecting, identifying, tracking and re-recognizing the target of attention of the monitoring video to achieve the key information extracted from the massive information and associates the same pedestrian in the video shot by the non-overlapping sight camera. Under the surveillance system without overlapped viewpoints, the difficulty of pedestrian re-recognition has been increased because of these complicated situations, such as perspective [10], illumination [18], complicated backgrounds [19] and obstructions [20], occlusion [4], [21], pose estimations [22], low resolution [5], [23] of the images from the video.

In an early work on person re-identification, Lin *et al.* [10] proposed deep features, which are automatically extracted by Convolution Neural Network (CNN) and then they described the features of image regions through the semantic image representation. Recently, the person re-identification is regarded as a classification problem by Fan *et al.* [24], and the difference between the features of the image and the sample is calculated to determine whether it belongs to the same pedestrian or not. High-level abstracted features [25], [22] deliver great performance in recognition. Yang *et al.* [26] and Zhao *et al.* [27] found that using multiscale convolution layer blocks and triplets of objects in unconstrained environments can obtain more discriminative features. In 2016, Varior et al. presented a new Siamese Long Short-Term Memory (LSTM) method that remembers the spatial connections of the image, processes image regions sequentially and enhances the degree of discrimination of local feature representation [28].

Feature extraction is a core content of the technology. A large number of pedestrian re-identification algorithms based on feature representation have emerged in recent years [8], [10], [29], [30]. In the surveillance video, it is easy to extract and express the features of pedestrians that will not change too much in a specific period of time [31]. According to literature, it is always difficult to choose a suitable and best feature extraction method for features extraction.

Recent studies [32] show that the images can be divided into several regions, and a variety of different low-level visual features are extracted from each region. For example, Liu *et al.* [33] utilized features of Local Maximal Occurrence (LMO) and a sliding window is used to describe the local features of a pedestrian image and the invariance of the angle of view. Nanda *et al.* [34] proposed a person recognition algorithm based on local feature classification in the same year. Moreover, The attribute features are proposed to determine whether the two images belong to same person by more semantic properties [35]–[37]. For instance, S.Gong adopted multiple semantics to describe person and used the support vector machine (SVM) to obtain the above-related attributes [38]. Different semantic attributes based on the importance of attributes fusion with low-level visual features are used to describe person images. It is considered that the attributes are usually the same in different monitoring videos, so the accuracy can be improved through the middle layer of attributes [39].

Through the process of attribute learning, the combination of each level of the image features is incorporated [6], [40]. Moreover, attribute learning has an irreplaceable role to describing person features in this field. By adding attribute layers between persons' features and persons' classifications, person re-identification can be improved with a better semantic expression [41]–[43].

III. METHOD

A. MULTI-CAMERAS LOGICAL TOPOLOGY INFERENCE

In our model, the Time Delayed Mutual Information (TDMI) is adopted to infer multi-cameras logical topology, which has showed its effectiveness and simplicity [16].The input to TDMI stands for the time series $S_{i,j} = (S_{i,j,1}, \ldots, S_{i,j,t}, \ldots)$, where $S_{i,j,t}$ refers to the feature code within the j-th region of i-th camera view at time t. TDMI was put forward to have a better understanding of the dependencies among the patterns of activity detected by cameras network. In form, the regional activity of two arbitrary patterns in two camera views are appointed to stand for two-time series by utilizing any types of encoding schemes mentioned above and denoted as $S_1(t)$ and $S_2(t)$. The TDMI of $S_1(t)$ and $S_2(t+\tau)$ is shown as follows:

$$I(s_1(t); s_2(t+\tau)) = \sum_{i=1}^{Ms_1} \sum_{j=1}^{Ms_2} Ps_1s_2 \log 2 \frac{Ps_1s_2(i,j)}{Ps_1(i)Ps_2(j)}$$

where Ms_1 and Ms_2 refer to the total number of bins of both $S_1(t)$ and $S_2(t+\tau)$, while $Ps_1(.)$ and $Ps_2(.)$ stand for distribution of marginal probability of $S_1(t)$ and $S_2(t+\tau)$ respectively; $Ps_1s_2(.)$ represents their distribution of joint probability. TDMI stands for measurement of symmetrical dependency between two-time series and $I(S_1(t); S_2(t+\tau)) \geq 0$, with the equality holding if and only if $S_1(t)$ and $S_2(t)$ are isolated with each other.

Calculating TDMI with various time delays $-T \leq \tau \leq T$ provides us with a TDMI function $I_{S_1S_2}(\tau)$ between the two regions:

$$Is_1s_2 = (I(s_1(t); s_2(t-\tau)); \ldots; I(s_1(t); s_2(t+T)))$$

The core constituents of the proposed approach are displayed in Fig.1. Because of the disjoint camera views within a camera network (Fig. 1(a)), we first extract the local Spatio-temporal patterns as the data of time series from every camera view (Fig. 1(b)). Then, the Cross Canonical Correlation Analysis (xCCA) is adopted to conclude the inter-region correlations that are delayed in time (Fig.1(c)). Afterward, we use correlation matrix to describe regional activity correlations (Fig.1(d)). The process of scene decomposition based on activity and xCCA is called training of the study. At last, camera's topology(Fig.1(e)) can be inferred by correlations of regional activity, which will offer a searching order for the re-identification of patient (Fig.1(f)).

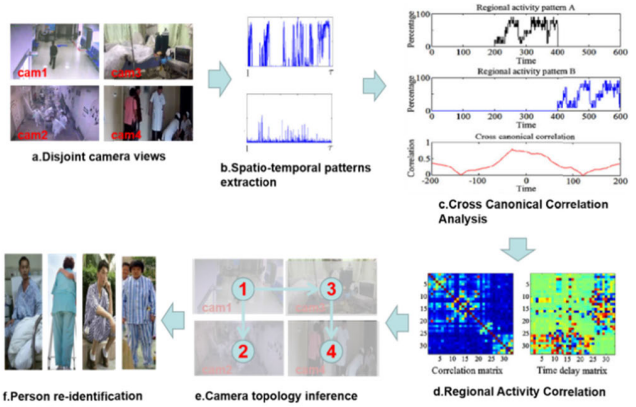

FIGURE 1. A diagram illustrating our multi-cameras logical topology inference.

B. FEATURE EXTRACTION FOR PATIENT RE-IDENTIFICATION

During the phase of re-identification of patients, Deep Convolutional Neural Network (DCNN) is adopted for feature extraction to make the attribute model that is driven by data that work for classification. Inspired by the model of [44], a DCNN, including an input layer, 4 hidden layers and a full convolution output layer (as shown in Fig.2) is designed for feature extraction. In the first hidden layer, there are 25 output feature maps, with the convolutional kernel size of 5×5. The second, third and fourth hidden layers have 50 output feature maps and the kernel size is still 5×5. Max-Pooling window is designed to be 2×2, and maps keep the size after max pooling, but the pixels are set to 0 except the max pixel in the window. Maps size changes to be half after pooling operation,

with those pixels, value 0 being removed. The last layer is the output layer, transforming all feature maps into a vector. We apply DCNN into the unsupervised learning features from one original image, which is the basis of target recognition of attributes.

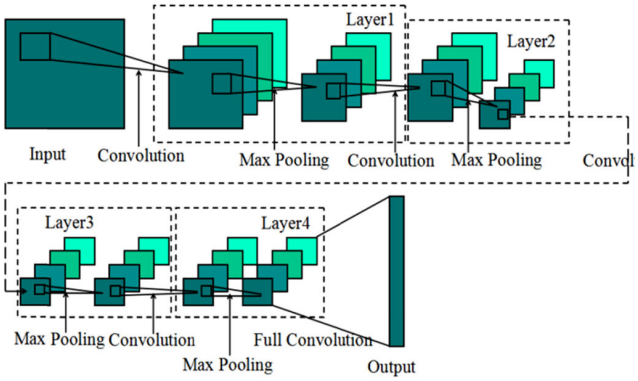

FIGURE 2. Framework of the deep convolutional neural network we used.

C. CLASSIFICATION BASED ON ATTRIBUTE

As we know, there are two kinds of modes for classification.One mode is based on the generative model,and the other mode is according to discriminative model.Based on the attribute, the classification is adopted inorder to handle the issue of re-identification of person. Thefollowing displays the framework of classification based onattributes:

If there is a set $\mathrm{X} = X_{train} \bigcup X_{test}$ and for each samplefrom the X_{train}, the class $\mathrm{y} \in Y$ and attributes representation $\mathrm{a} \in A$ is known, then we can learn a non-trivialclassifier $\mathrm{f} : x \xrightarrow{a} y$ to label the class of sample x fromtest set with the help of attributes, where Y is the set of classes, and A is the set of attributes.

A classifier for attribute a, trained by labeling some images of all classes for which $a_i = 1$ as positive andthe rest as negative training examples, can provide anestimate of the posterior probability $P(ai\,|x)$ of that attributebeing present in image. Mutual independence yields $\mathrm{P}(a\,|x) = \prod_{i=1}^{M} p(a_i\,|x)$ for multiple attributes. The posteriorprobability of class y is shown in image x, which is acquired bymarginalizing all probable associations among attributes,by Bayes rule [45]:

$$\begin{aligned} \mathrm{P}(y\,|x) &= \sum_{a\in\{0,1\}M} P(y\,|a)P(a\,|x) \\ &= \sum_{a\in\{0,1\}M} \frac{p(a\,|y)p(y)}{p(a)} p(a\,|x) \end{aligned} \tag{1}$$

where $p(a\,|y) = \prod_{i=1}^{M} p(a_i\,|y) = \prod_{i=1}^{M}\prod_{c=1}^{C} p(a_i\,|y_c)1^B$ M refers to attributes' number, C stands for classes'number, $\mathrm{B} = \{a_i\,|a_i \in A,\ where\ a_i = 1\ \ and\ \ y_{c=1}\}$. Fig.3 shows an example of class-attribute association matrix B. Taking into account that both $p(y)$ and $p(a)$ represent theevidence

y / a	y_1=1 $y_{i\neq1}$=0	y_2=1 $y_{i\neq2}$=0	⋯	y_c=1 $y_{i\neq c}$=0
a_1=	0	1	⋯	0
a_2=	1	1	⋯	0
⋯	⋯	⋯	⋯	⋯
a_M=	1	0	⋯	1

FIGURE 3. Class-attribute association matrix.

probability not affected by samples in test, theoutput of classifying $\widehat{y} = \max_{y} p(y\,|x)$ only depends on $p(a\,|y)$ and $p(a\,|x)$. Therefore, how to obtain the attributes of thepictures of set B is the key procedure for the frameworkof classification based on attribute. Following twoparts will provide an approach to handle the issue. Theclassification based on attribute is demonstrated in Fig. 4. The output of the DCNN model can be employed as theinput of the attribute-based classification model.

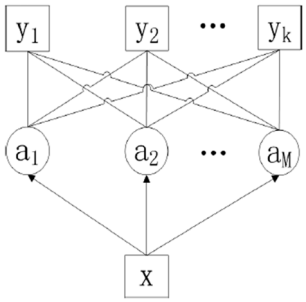

FIGURE 4. Classification model based on attributes.

D. ATTRIBUTE LEARNING WITH STRUCTURAL OUTPUT

Many patients' attributes are closely related toa local area or part of the patients, for example, "a patient with blackshoes". That is to say, the classification of attributesshould output with structural coordinate. Consider aset of input images $\{x_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}\} \subset \mathrm{x}$ and their boundary $\mathrm{O} = \{o_1, o_2, \ldots, o_n\}$ of one attribute $a \in A$, wehope to investigate a mapping $\mathrm{g} : X \rightarrow \mathrm{O}$ that we are ableto annotate in an automatic way whether the test imagehas the attribute a_i or not. We deliberate over the casethat the output space composes of one label to indicateif an attribute is demonstrated, and a vector indicates top,bottom, left and the right of bounding box in area of theattribute: $\mathrm{O} \equiv \{\alpha, y, l, b, r\,|\alpha \in \{0, 1\}, (t, l, b, r) \in R^4\}$. If $\alpha = 0$ the coordinate vector (t, l, b, r) is overlooked. Themapping in structured framework of learning is shown as:

$$\hat{\mathrm{O}} = \arg\max o\{w^T \phi(x, o)\} \tag{2}$$

where $\varphi(x, o)$ is a joint kernel map and $\mathrm{x}^T\varphi(x, o)$ stands for a discriminative function which is expected toprovide great value for pairs (x, o), which are matchedas well. In order to train that discriminator, Cutting PlaneAlgorithm [46] is used

to obtain the best w and ξ of a given attribute, as described in Eq(3).

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + \xi$$
$$s.t.\ \xi \geq 0, \forall o_i' \in o \backslash oi \frac{1}{n}\sum_{i=1}^{n}\left[w^T\varphi(x_i, o_i) - w^T\varphi(x_i, o_i')\right]$$
$$\geq \frac{1}{n}\sum_{i=1}^{n}\left[\Delta(oi, o_i')\right] - \xi \tag{3}$$

The Cutting Plane Algorithm [47] is employed toestimate w and ξ, where the model is trained by utilizingconstraints' subsets, besides, the brand new constraintsare added by finding o which can maximize the rightside of hand of Eq(3). The alternation does not repeatuntil the convergence with a small set of constraints ismade comparison with the size of O. As a result, we areable to optimize the discriminative function in an effectivemannerbased on a choice of loss $\Delta(oi, o)$ and the kernel $k((x, o), (x', o'))$.

The choice of loss function $\Delta(oi, o)$ is supposed toreflect the quantity that measures how excellently thesystem performs. We select the loss that is constructed by the measurement of area overlap and it is described as follow:

$$\Delta(oi, o) = \begin{cases} 1 - \dfrac{Area(o_i \cap o)}{Area(oi \cap o)}, & if\ O_{i\alpha} = O_\alpha = 1 \\ \dfrac{1}{2}, & otherwise \end{cases} \tag{4}$$

where $o_{i\alpha} \in \{0, 1\}$ indicates whether the attribute a is absent or present in the imagearea. $\Delta(oi, o)$ possesses the expectedperformance, which is equal to zero wherethe offered bounding boxes are identical and is equalto 1 when being disjoint. The formulation (4) is attractive in that it scales smoothly with the degree of overlap between the solutions, which is important to allow the learning process to utilize partial detection for training.Therefore, if the label reaches an agreement, the loss is 0; if these labels disagree, the loss is 1, which creates the usual SVM concept of margin.The setup performs automatically with an approachof the maximum margin of two conditions that are significantfor localization. Firstly, for images with attributes that will beobserved and the localized region is supposed to acquire the topscore of all applicable boxes. Secondly, for images withoutattributes and no box is expected to obtain a top score.

E. ATTRIBUTE OBTAINING

1) Numerous methods of pattern recognition can be utilized for the re-identification of person. Traditional methodsadopt the visual features of low level to portray people. Nevertheless, one individual appearance possesses the high-levelsemantic features. The gap between low-level features and the high-level categories make there-identification task of a person quite hard. As a result of intermediary of semantic gap, the assisted attributes canplay a significant role in the re-identification of person [48], [49].

2) For the appropriate attribute, a question is arise that how to acquire people's appropriate attribute? In general, two ways can tackle the problem. The first way is the clothing benchmark. For instance Deepfashion2 [50] obtained a lot of clothing attributes by analyzing the pose estimation, detection, re-identification. However, some attributes ofappearance only offer a good semantic explanation to users but they are not useful to distinguish the category. The attributessearched by us for the re-identification of person shouldtake both categories to distinguish ability and the semanticexplanation into account. There is a method to obtain attributes named ' manual annotation'. Layne *et al.* [51] organizedpeople to write down their attributes and annotate the binaryvalues. Further, these attributes are utilized to re-identifypeople.

Although the attributes of people listed by experts can be acquired, butstill we cannot ensure that those attributes will have enough effectiveness in the classification when the feature space of an image is divided. Contrarily, even if we find out the margins of various feature spaces that were not helpful to identify people, we still cannot guarantee if the hyper-planes are nameable or semantic by human.

Fig. 5. Displays a summary of our method to exploit both discriminative and understandable attributes of people based on data driving. First, a hyperplane is produced in the feature spaces based on present set of attributes in order to reduce the confusion of various categories. Second, a discriminative model is adopted to evaluate if the candidate assumption on attribute is nameable. Each acceptedattributes will be added into training dataset of the model. When a candidate assumption on attribute is evaluated to be unnamable, it will be abandoned and another assumption will be estimated. If the assumption is nameable, it will be visualized by showing samples that are related to the attribute. Finally, annotator based on the visualized assumption will name the candidate attribute. When the naming succeeds, a new attribute a_j will be acquired and is added to the current attribute set, namely, $A_{t+1} = [A_t; a_j]$. In the meantime, the model with nameability will be renovated with a newly added training sample. Afterward, the system will produce a new candidate assumption on attribute to estimate. If there are sufficient numbers of named attributes to recognize the categories, the loop will end.

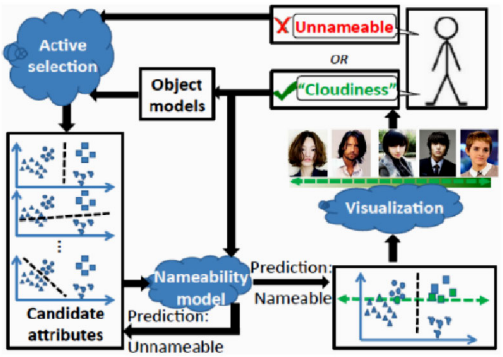

FIGURE 5. Overview of our attribute obtained approach.

F. PARALLELIZED IMPLEMENTATIONS OF DISTRIBUTED DEEP ATTRIBUTE LEARNING MODEL

Distributed Deep Attribute Learning model (DDAL) based on DCNN is exhibited to be proficient in the task of classification and it possesses the ability to operate on the raw pixel input without the help of special features of design, which is quite appealing. Nevertheless, it is significantly slow at inference time. As a result, we adopt the parallelized implementations to make the recognition faster. The core thought of the parallelized implementations of our model is the attribute learning and the distributed parameter manipulation. For the distributed parameter manipulation(Fig. 6.), system's coordinates will send commands to every replica machine of the model. Then, the feedback offered will be stored in the server of parameter. In addition, some history cache to optimize the algorithm is sent to the shard of server, on which it was calculated [52], so as to prevent sending every parameter to the iteration of central server. In the distributed model, these samples are distributed to various machines and then the output results of these machines will be collected into the central server of parameter. In the traditional distribution model, the machines that run at the slowest speed are always an obstacle to the model. The entire system cannot but wait for the slowest machine, which makes the model difficult to handle those major issues. In order to tackle this issue, we adopt the following method: only a small portion of data can be sent to every model replica when one machine is available and new data will be added into it via coordinator. In other words, if one machine runs faster, it is expected to handle more data. While handling the final batch, the coordinator will send copies of the batch to each machine. When one of them complete the first, the final result of it will be accepted, which is similar to the employment of ''backup tasks'' of ''MapReduce''scheme [53]. In the operation process of parallelized system, the data with the relation is fetched in advance to the same machine. The fetching of data with the supportive data affinity through the assignment of data's sequential portions to the same worker allows the system to run in a smooth way.

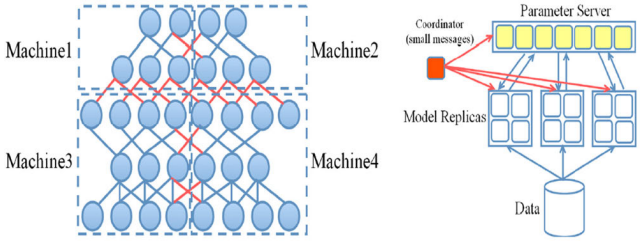

FIGURE 6. Distributed parameter manipulation of our model.

Another parallelized implementation in model is the distributed learning of attributes. The learning model of attributes is segmented through multiple machines as learning tasks of these attributes are assigned to various machines. During the process of both inference and training, the model's coordinator parallelizes every machine to calculate in an automatic way and manages the synchronization, communication, and information that are transferred among these machines.

IV. EXPERIMENTS

A. FEATURE AND DATASET SELECTION

Three challenging datasets are selected: AHU, i-LIDS as well as CUHK dataset [54], [55], in order to verify our model.

AHU The affiliated hospital of our university dataset contains fix views from 22 disjoint cameras installed at the hospital (see Fig.7). The length of videois 177 hours in length with 0.7 fps of frame rate which is taken from these cameras. The size of every frame is 320×230. We divide the dataset into 10 subsets and every subset has 5000 frames per camera. We use two subsets as the validation data. As for the residual weight subsets, 500 frames/camera from each subsetare utilized to train and test the rest.

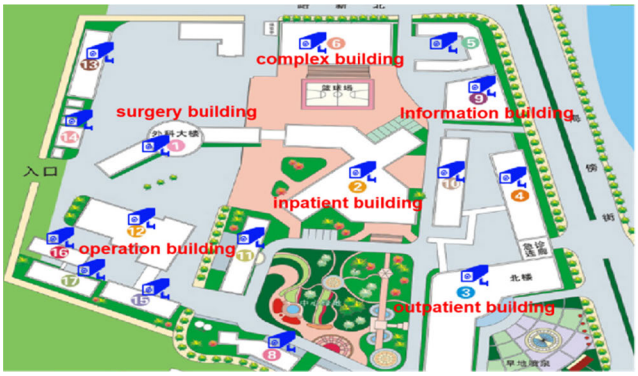

FIGURE 7. The topology map of multi-cameras.

i-LIDS[56] is acquired with 721×577 resolution and 25 fps by five static and synchronized cameras which are disjoint and installed at a hectic airport. The pair of cameras that are selected for the dataset possesses a gap of shorter time, as compared with the AHU dataset. Nevertheless, different from AHU dataset, the two views chosen of i-LIDS dataset can obtain quite diverse view fields. Like,camera 1 possesses a close view field while camera 2 has a wider zone view field comparatively. The remarkable difference in the view field enhances the difficulty to correlate and match visual features through views.

CUHK is an occlusion dataset. It contains 1063 frames from the datasets of Caltech [56], ETHZ [57], TUD-Brussels [58], INRIA [59], Caviar [60]. Some frames of this dataset contain occluded pedestrian while some do not.

In feature selection process, we employ DDAL to acquire these features based on DCNN, which is trained before by the minimization of reconstruction error. Fig. 8.is the visualization of weights learned in the first convolution layer of DCNN. Fig. 9.demonstrates the activation of DDAL based on DCNN and takes it as an example of input. Every panel displays the normalization in the convolutional layer, follow by poolinglayers, then the 1×1 convolutional one is shown and last layer is fully-connected.

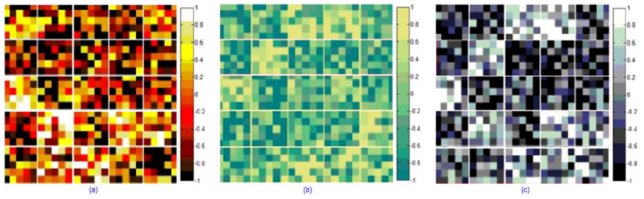

FIGURE 8. Visualization of weights learned in the first convolution layer. From (a) to (c) are filters for red, greenand blue. They are 3 color channels of the RGB person image.

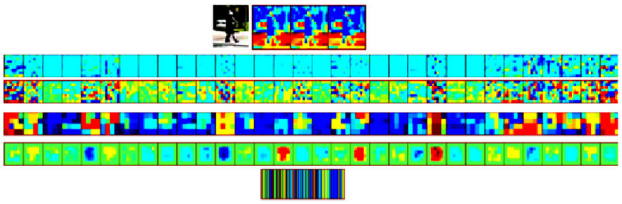

FIGURE 9. Activation of DDAL for an input example.

B. MULTI-CAMERA TOPOLOGY INFERENCE

In experiments related to topology inference, TDMI model is compared with CCA, as well as structure learning of MCMC Bayesian network for the learning correction of regional activity. Fig.10 shows the affinity matrices of regional activity. According to TDMI affinity matrix in Fig. 10(c), it accurately founds the correlation between both regions through camera's views. By contrast, affinity matrix of CCA tends to show these regions within the same camera view, which reveal the high correlations only. Similarly, the structure learning of MCMC Bayesian network implies few and inaccurate correlations because of enormous missing detection.

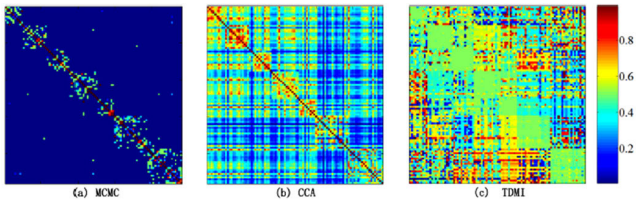

FIGURE 10. Regional activity affinity matrices.

Fig.11 shows the camera topology that is yielded by various methods. Based on the final results, we can observe that the closest topology TDMI is the actual topology. It is quite common that TDMI remarkably exceeded the structure learning of MCMC Bayesian network. In MCMC, the structure that uses the structure learning of Bayesian network only discloses dependencies of zero-order temporal, i.e. co-occurrence relations among activities. Therefore, more complicated correlationscannot be handled which are common within a multi-camera scene. The reason why TDMI exceeded CCA is that it owns the capability to capture these potential mutual patterns of both time series of regional activity with the means of projecting them on an optimal subspace, which is significant to analyze the hectic public

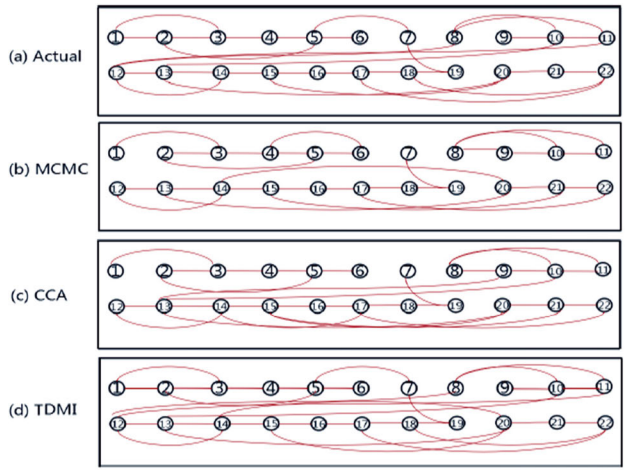

FIGURE 11. Our multi-cameras logical topology inference compared to other methods.

spaces where important variations for correlative activities exist with various views. These views are obtained with the help of different angles of the camera and uncertain time delay of activity among them.

C. ATTRIBUTE OBTAINING & Prediction

In the experiment, we set up a name-ability manifoldwith low dimension utilizing the instances of parameters ofSVM hyperplane (weight bias and vector) which conformto the real nameable attributes. In particular, the nameableattributes are modeled with a mixture of probabilisticprincipal component analyzers (MPPCA). We use the visualizedattribute to help to exploit the attributes. In orderto name the attribute assumptions, we show itwith some visual samples, particularly sample's variationbetween hyperplane's both sides. A majority of instancesare taken into account except for some samples of an outlier. Thesample space is classified by attribute that is divided into severalsubspaces, in which three or four instances are chosen asthe representatives respectively. Fig.12 displays an instanceof visualized and named attribute. Fig. 13 demonstratesthe attributes that are exploited by our model. Besides,these individuals from training dataset are labeled with theattributes.

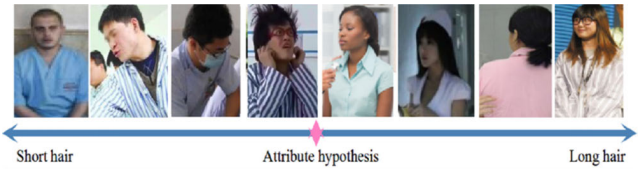

FIGURE 12. Visualizing an attribute hypothesis.

In order to make a mapping between class and attribute, image attributes prediction is a crucial routine. At the startof process, the responses of people tested will determinethe association strength of real value between picture andattributes on average. The threshold value helps to obtain the

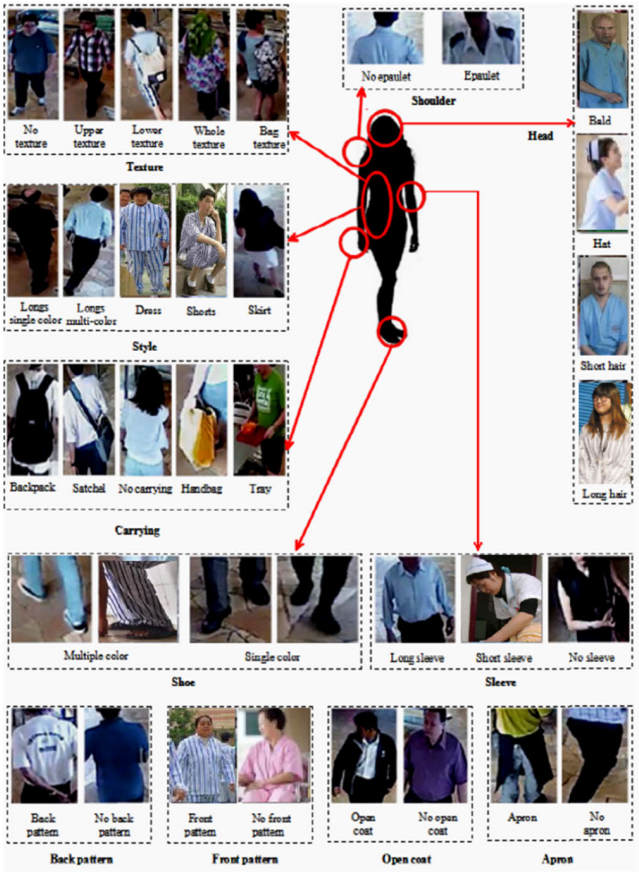

FIGURE 13. Attributes obtained.

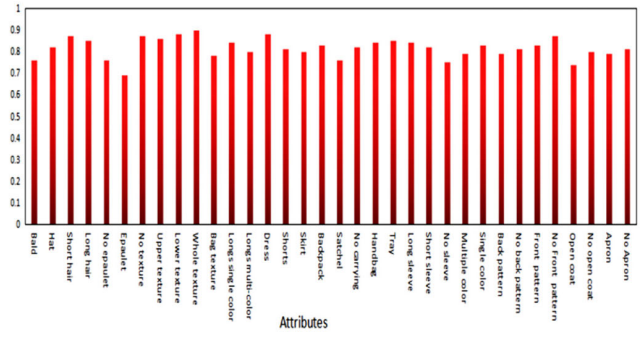

FIGURE 15. Quality of individual attribute predictors.

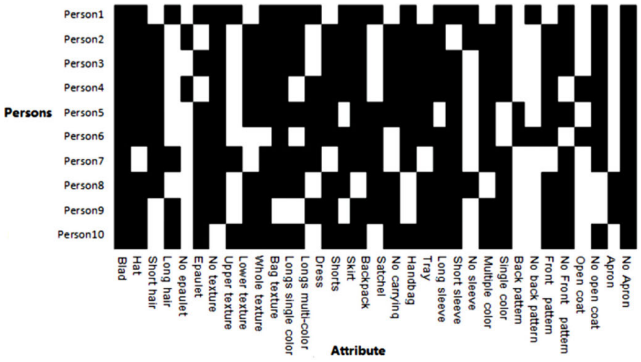

FIGURE 16. Visualization of attribute-class matrix.

binarymatrices of sample attribute at the mean of the overall matrix. In addition, we conduct an analysis about the influence of asingle attribute on the accuracy of prediction. The imagesfrom the attributes which are labeled by a test set of human are utilized,in order to test the quality of the predictors for each attribute. There is an instance of the prediction and location for partattributes in Fig.14 and Fig. 15, which illustrates the accuracy of prediction of every attribute. It is quite easy to predict some attributes while some attributes are very difficult to predict because the ability to identify a specific attribute is associated with the universality. Those attributes that can be observed easily are much easier to make recognition than these ambiguous attributes. In Fig.16, a part of attribute class matrix is visualized to disclose the correlation between every category and attribute. On account of the attribute predictor and the attribute-class matrix, the task of classification of person category will be fulfilled [43].

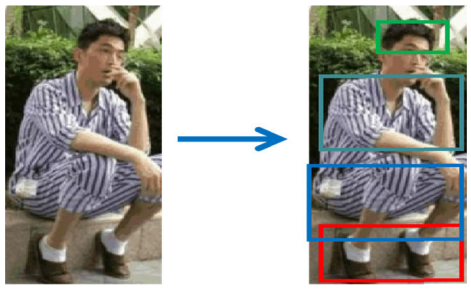

FIGURE 14. Some attributes' prediction and location.

D. MATCHING COMPARISON OF PERSON RE-IDENTIFICATION

At first, with the classification and evaluation protocol that is discussed above, every matched sample from the frames taken by one camera is matched with each sample from the camera by the order of logical topology. Then, the accurately matched rank can be obtained. In particular, the recognition rate of $rank - k$ signifies the matches' expectation at the rank k. The CMC curve represents the accumulative values of the recognition rate at every rank. Later, the set of matched sample is switched with the match, and the mean value of two rounds of the curves is used as the final result of a test. The average evaluation is a result of repeated ten tests is considered as a stabilized statistical result. We repeat these tests for ten times and their average value is considered as a stabilized statistical result.

As DF, ELF, bLDFV, SDALF, SDC-knn, SDCocsvm, eSDC-knn and eSDC-OCSVM [61], [62] have published the recognition accuracy therefore, they are suitable for comparison. In the experiment, some test data and training data are assigned in these diverse models that are compared. The final results of these comparisons are demonstrated in Fig. 17. It is

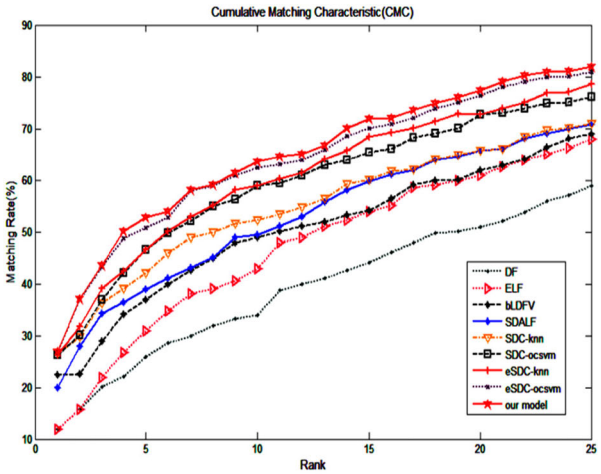

FIGURE 17. Rate of cumulative matching characteristic on AHU dataset.

proved that our approach is superior to all the benchmarking approaches and can give a good semantic explanation. Especially, the matched rate of our model is approximately 27% at rank-1 and about 77% at rank-25.

Besides, the ''zero-shot'' scene of the re-identification system is still allowed to be operated if there is a relevant semantic description provided by the user. The accuracy of our model reached approximately 26.5% in the final re-identification of person on AHU dataset and about 22.1% on i-LIDS dataset with the zero samples of training, respectively. In addition, the model is compared with six person re-identification methods, including four methods based on attributes, OAR [63], SDCknn, Umeda's method [64], and SDCocsvm [62]. Table 1 describes the accuracy of six models. We can see that the proposed model obtained the highest classification accuracy for both datasets as compared to other methods. In [65], authors verified their AIR and weighed AIR models on i-LIDS dataset, and the re-identification rate for transfer learning of zero-shot are 11.5%

TABLE 1. Comparison of re-identification accuracy with zero-shot training on AHU and i-LIDS.

Methods	Datasets	
	AHU	I-LIDS
ORA	24.1	-
Umedaet.al.	18.0	-
SDC_knn	25.3	-
SDC_ocsvm	25.7	-
AIR	-	11.5
W.AIR	-	16.5
Our model	**26.5**	**22.1**

and 16.5%, which is lower than our model. Experiments prove that our model significantly improves the degree of accuracy and enhances the semantic representation of re-identification of person based on attributes and deep learning.

Another challenging task is how to handle the re-identification of patients with partial occlusion. Occlusions are commonplace in a crowded public space with background clutters, static barriers such as the wall and pillar and other things in the scene etc. Occasionally, some changes can be brought about in a deliberate way. For instance, if criminal activity has been reported by media,the police of CCTV personnel who inspect surveillance video usually find that only some parts of the criminal suspect's body are visible in the scene, for the suspect attempts to hide the appearance deliberately.The purpose of the case is to find out the person appeared in other surveillance videos with the complete appearance. In our experiment based on the CUHK occlusion dataset, we compare our model with HOG+SVM and LatSVM-V2 [66] in the case of samples with occlusion and without occlusion. The experimental results are shown in Fig. 18.

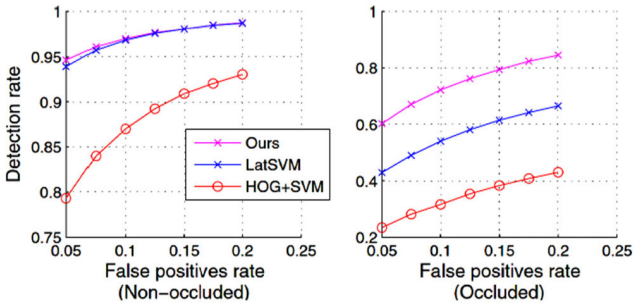

FIGURE 18. Experimental results on CUHK occlusion dataset.

E. INCREASING THE SPEED OF MODEL PARALLELISM

We accomplished our model based on Caffe framework with Linux version [67]. We conduct the program on a cluster thatis composed of about 15 machines. Every machine possesses a GTX780 GPU with a memory of 2G, a gigabit Ethernet link and two160GB IDE disks. Moreover, about 100 Gbps of aggregated bandwidth that is obtainable at network root is arranged for machines. To have an understanding of the performance of the model parallelism, we make measurement on the average time to train partitions (machines) utilized in our model. Comparison of the meantime in training between usage of one single machine only and N machines is displayed in Fig.19.

Fig.19 indicates that with eight machines, the model with a moderate size, such as our 34-attribute model operates 2.3 times faster than one single machine. If there are more than eight machines, system's overall property will start to decrease, because more partitioning can decrease the working capability of machines. Moreover, network overhead plays the role of time-consuming element of the system. By comparison, when there are more attributes or parameters in

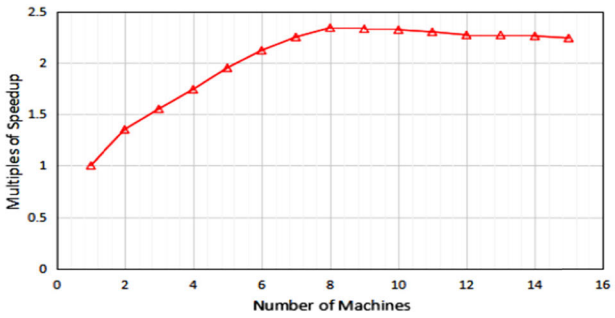

FIGURE 19. Increase in speed of model parallelism with the increase in number of machines.

a model then utilizing more machines will be preferred to conduct parallel computing.

V. ANALYSIS

A. TIME DELAYED MUTUAL INFORMATION & MULTI-CAMERA TOPOLOGY INFERENCE

In the experiment section, we use TDMI to inference the multi-camera logical topology. Compared with CCA and MCMC, TDMI can accurately found the correlation between both regions through camera's views. While in CCA model, affinity matrix of CCA tends to show these regions within the same camera view, which reveal the high correlations only. In MCMC, the structure that uses the structure learning of Bayesian network only discloses dependencies of zero-order temporal, i.e. co-occurrence relations among activities. Therefore, correlations are more complicated and cannot be handled in a multi-camera scene.

Above all, the result shows that TDMI owns the capability to capture these potential mutual patterns of time series of regional activity with the means of projecting them on an optimal subspace, which is significant to analyze the hectic public spaces where important variations for correlative activities exist with various views. These views are obtained with the help of different angles of the camera and uncertain time delay of activity among them.

B. STRUCTURAL OUTPUT ATTRIBUTE LEARNING & PERSON RE-IDENTIFICATION

In this experiment, we made a comparison between our learning method of attribute with several learning methods without attribute, such as SDF-knn and ELF. We can observe from Fig. 14 that our approach can offer a structural output of attributes which certainly provides a great semantic explanation and has a capability to recognize these new categories from these descriptions of plain text.

Especially, structural output of attributes is helpful to increase the precision of attribute classifier and improve the robustness of the model, such as the ability to against the very low resolution, occlusions and pose, viewpoint and illumination changes. The reason for not using the class labels of samples to train a category classifier for person (patients) re-identification is that samples with category labels are more difficult to collect as compared to samples with attribute labels. In addition, the number of categories is much more than the number of attributes. Therefore, we employ an attribute learning model to address person re-identification task. Besides, attributes can play a crucial role not only in the prediction of sample learning but also in the learning of zero-shot. In other words, the attributes can be acquired and learned without understanding person's class labels. This observation supports the argument that why attributes are reliable to use in transfer learning.

C. DEEP LEARNING & PARALLELIZED MODEL

In our model, we employ Deep Convolutional Neural Network(DCNN) to deal with feature extractions for the attribute model. We can utilize a large number of samples without labels to train the deep network in an unsupervised way and just need a small number of labeled samples to fine-tune the weights of the trained deep network in the training stage. In short, the proposed model is a data-driven classifier which has showed more generalized performance as compared to many shallow learning models, such as SVM.

Otherwise, compared with traditional serial models, we use parallelized implementations to improve our model. It contains data parallelism (divide the image data into several pieces), model parallelism (employ several machines to complete the task cooperatively) and attributes parallelism (different learning model of attributes can be segmented through different machines). According to the result in Fig.19. the model with a moderate size, such as our 34-attribute model, operates 2.3 times faster than one single machine. It means that these parallelized implementations can accelerate DCNN serial models processing speed obviously.

VI. CONCLUSION

In this paper, we have discussed a data-driven way to address how to use anIoT with distributed cameras to re-identify patients in a private space such as health care hospital and asylum to prevent the risks of violence from patients and sometimes-abusive behavior of staff that leads to deaths of patients. We employ a Time-Delayed Mutual Information (TDMI) model to address multi-cameras logical topology inference thathas been given a searching order for our task. Instead of training for the recognition of a specific category of patients directly based on the manually designed features, a series of visual attributes with structural output are extracted from a given set of images, which are mined consideringboth human understandable and discriminative demand.Moreover, we adopted Deep Convolutional Neural Network (DCNN) tohandle the lack of labeled samples in task of person re-identification and to produce features. A large number of samples without labels have been utilized to train the deep network in an unsupervised way. However,small amounts of labeled samples are needed to fine-tune the weights of the trained deep network in the training stage. In short, the proposed model is a data-driven classifier which showed

more generalized performance as compared to many shallow learning models. Furthermore, the parallelized implementations, such as the attribute learning and distributed parameter manipulation are used to accelerate the model. The experimental results prove the efficiency of the proposed model for patients' re-identification along with a great semantic explanation.

REFERENCES

[1] A. Prati, C. Shan, and K. I.-K. Wang, "Sensors, vision and networks: From video surveillance to activity recognition and health monitoring," *J. Ambient Intell. Smart Environ.*, vol. 11, no. 1, pp. 5–22, 2019.

[2] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, vol. 46, pp. 147–170, Mar. 2019.

[3] K. T. Chui, P. Vasant, and R. W. Liu, "Smart city is a safe city: Information and communication technology–enhanced urban space monitoring and surveillance systems: The promise and limitations," in *Smart Cities: Issues and Challenges: Mapping Political, Social and Economic Risks and Threats*. London, U.K.: Elsevier, 2019, pp. 111–124.

[4] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, and Z. Zhu, "Robust face recognition via occlusion dictionary learning," *Pattern Recognit.*, vol. 47, no. 4, pp. 1559–1572, Apr. 2014.

[5] X. You, W. Ou, C. L. P. Chen, Q. Li, Z. Zhu, and Y. Tang, "Robust nonnegative patch alignment for dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2760–2774, Nov. 2015.

[6] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2018.

[7] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," 2019, *arXiv:1907.06670*. [Online]. Available: https://arxiv.org/abs/1907.06670

[8] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[9] L. Wu, Y. Wang, X. Li, and J. Gao, "What-and-where to match: Deep spatially multiplicative integration networks for person re-identification," *Pattern Recognit.*, vol. 76, no. 3, pp. 727–738, 2018.

[10] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[11] P. Zhu, H. Wang, and V. Saligrama, "Zero shot detection," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[12] Y.-J. Cho, S.-A. Kim, J.-H. Park, K. Lee, and K.-J. Yoon, "Joint person re-identification and camera network topology inference in multiple cameras," *Comput. Vis. Image Understand.*, vol. 180, pp. 34–46, Mar. 2019.

[13] S. Jain, G. Ananthanarayanan, J. Jiang, Y. Shu, and J. Gonzalez, "Scaling video analytics systems to large camera deployments," in *Proc. 20th Int. Workshop Mobile Comput. Syst. Appl.*, 2019, pp. 9–14.

[14] M. Wolf and J. Schlessman, "Distributed smart cameras and distributed computer vision," in *Handbook of Signal Processing Systems*. Cham, Switzerland: Springer, 2019, pp. 361–377.

[15] C.-T. Chu, J. Jung, Z. Liu, and R. Mahajan, "Secure and private tracking across multiple cameras," Google Patents 20 140 184 803 A1, Jul. 3, 2014.

[16] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *Int. J. Comput. Vis.*, vol. 90, no. 1, pp. 106–129, 2010.

[17] R. Prates and W. R. Schwartz, "Kernel cross-view collaborative representation based classification for person re-identification," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 304–315, Jan. 2019.

[18] A. Bhuiyan, B. Mirmahboub, A. Perina, and V. Murino, "Person re-identification using robust brightness transfer functions based on multiple detections," in *Proc. Int. Conf. Image Anal. Process.*, vol. 9280, 2015, pp. 449–459.

[19] D. Wu, S.-J. Zheng, W. Z. Bao, X.-P. Zhang, C.-A. Yuan, and D.-S. Huang, "A novel deep model with multi-loss and efficient training for person re-identification," *Neurocomputing*, vol. 324, pp. 69–75, Jan. 2019.

[20] S. Moosavi, "Deep person re-identification using supervised learning with ranking method," Univ. Texas San Antonio, San Antonio, TX, USA, Tech. Rep. 13879924, 2019.

[21] L. Du and H. Hu, "Nuclear norm based adapted occlusion dictionary learning for face recognition with occlusion and illumination changes," *Neurocomputing*, vol. 340, pp. 133–144, May 2019.

[22] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[23] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2000–2012, Aug. 2019.

[24] X. Fan, W. Jiang, H. Luo, and M. Fei, "SphereReID: Deep hypersphere manifold embedding for person re-identification," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 51–58, Apr. 2019.

[25] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8295–8302.

[26] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognit.*, vol. 86, pp. 143–155, Feb. 2019.

[27] C. Zhao, K. Chen, Z. Wei, Y. Chen, D. Miao, and W. Wang, "Multilevel triplet deep learning model for person re-identification," *Pattern Recognit. Lett.*, vol. 117, pp. 161–168, Jan. 2019.

[28] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 135–153.

[29] L. Wu, Y. Wang, L. Shao, and M. Wang, "3-D PersonVLAD: Learning deep global representations for video-based person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3347–3359, Nov. 2019.

[30] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 4321–4329.

[31] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, "Crowd counting in low-resolution crowded scenes using region-based deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 35317–35329, 2019.

[32] B. Shekar, S. K. Pavani, and A. Garg, "Distinguishing between stock keeping units using a physical dimension of a region depicted in an image," Google Patents 10 318 837 B2, Mar. 11, 2019.

[33] J. Liu, Z.-J. Zha, X. Chen, Z. Wang, and Y. Zhang, "Dense 3D-convolutional neural network for person re-identification in videos," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 15, no. 1s, 2019, Art. no. 8.

[34] A. Nanda, D. S. Chauhan, P. K. Sa, and S. Bakshi, "Illumination and scale invariant relevant visual features with hypergraph-based learning for multi-shot person re-identification," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 3885–3910, 2019.

[35] K. Cheng, Y. Zhan, and M. Qi, "AL-DDCNN: A distributed crossing semantic gap learning for person re-identification," *Concurrency Comput., Pract. Exper.*, vol. 29, no. 3, 2017, Art. no. e3766.

[36] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 598–607.

[37] V. Sharma and M. Zuliani, "System and method for person reidentification," Google Patents 10 318 721 B2, Mar. 30, 2019.

[38] S. S. Gong and T. M. Hospedales, "Visual data mining," Google Patents 10 282 616 B2, May 7, 2019.

[39] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8738–8745.

[40] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, "Looking beyond appearances: Synthetic training data for deep CNNs in re-identification," *Comput. Vis. Image Understand.*, vol. 167, pp. 50–62, Feb. 2018.

[41] K. Cheng, F. Xu, F. Tao, M. Qi, and M. Li, "Data-driven pedestrian re-identification based on hierarchical semantic representation," *Concurrency Comput., Pract. Exper.*, vol. 30, no. 23, 2018, Art. no. e4403.

[42] S. W. Bak and G. P. K. Carr, "One shot color calibrated metric learning for object re-identification," Google Patents 10 331 968 B2, Jun. 25, 2019.

[43] K. Cheng and X. Tan, "Sparse representations based attribute learning for flower classification," *Neurocomputing*, vol. 145, pp. 416–426, Dec. 2014.

[44] F. Li, H. Qiao, and B. Zhang, "Discriminatively boosted image clustering with fully convolutional auto-encoders," *Pattern Recognit.*, vol. 83, pp. 161–173, Nov. 2018.

[45] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where—And why? Semantic relatedness for knowledge transfer," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 910–917.
[46] Y. Yao, W. Yang, P. Huang, Q. Wang, Y. Cai, and Z. Tang, "Exploiting textual and visual features for image categorization," *Pattern Recognit. Lett.*, vol. 117, pp. 140–145, Jan. 2019.
[47] J. E. Kelley, Jr., "The cutting-plane method for solving convex programs," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 4, pp. 703–712, 1960.
[48] Z. Chen, A. Li, and Y. Wang, "A temporal attentive approach for video-based pedestrian attribute recognition," 2019, *arXiv:1901.05742*. [Online]. Available: https://arxiv.org/abs/1901.05742
[49] B. Xu, J. Liu, X. Hou, K. Sun, and G. Qiu, "Cross domain person re-identification with large scale attribute annotated datasets," *IEEE Access*, vol. 7, pp. 21623–21634, 2019.
[50] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5337–5345.
[51] R. Layne, T. M. Hospedales, and S. Gong, "Towards person identification and re-identification with attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 402–412.
[52] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, Dec. 2019.
[53] D. Dessì, G. Fenu, M. Marras, and D. R. Recupero, "Bridging learning analytics and Cognitive Computing for Big Data classification in micro-learning video collections," *Comput. Hum. Behav.*, vol. 92, pp. 468–477, Mar. 2019.
[54] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.
[55] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3258–3265.
[56] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 737–753.
[57] A. Ess, K. Schindler, B. Leibe, and L. Van Gool, "Object detection and tracking for autonomous navigation in dynamic environments," *Int. J. Robot. Res.*, vol. 29, no. 14, pp. 1707–1725, 2010.
[58] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 794–801.
[59] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.
[60] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. BMVC*, 2011, vol. 1, no. 2, p. 6.
[61] M. Farenzena, L. Bazzani, and A. Perina, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.
[62] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.
[63] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 93–117.
[64] T. Umeda, Y. Sun, G. Irie, K. Sudo, and T. Kinebuchi, "Attribute discovery for person re-identification," in *Proc. Int. Conf. Multimedia Modeling*, 2016, pp. 268–276.
[65] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proc. BMVC*, 2012, vol. 2, no. 3, p. 8.
[66] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[67] L. Gu and H. Li, "Memory or time: Performance evaluation for iterative operation on Hadoop and spark," in *Proc. IEEE 10th Int. Conf. High Perform. Comput. Commun. IEEE Int. Conf. Embedded Ubiquitous Comput.*, Nov. 2013, pp. 721–727.

KEYANG CHENG received the Ph.D. degree from the School of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, in 2015. He was a Postdoctoral Researcher with the University of Warwick, U.K., in 2016. He is currently an Associate Professor with the School of Computer Science and Telecommunications Engineering, Jiangsu University. He has coauthored more than 30 journal and conference papers. His current research interests include in the areas of pattern recognition, computational intelligence, and computer vision. He is a member of CCF.

MUHAMMAD SADDAM KHOKHAR received the M.S degree in computer science and information technology from N.E.D University, in 2016. He is currently pursuing the Ph.D. degree with the School of Computer Science and Telecommunications Engineering, Jiangsu University. He has a vast academic, technical, and professional experience. He supervised more than 30 academic FYP Research projects. His current research interests include computational intelligence, pattern recognition, deep learning, and computer vision.

QING LIU is currently pursuing the master's degree in computer application with Jiangsu University. His research interests include computer vision and pattern recognition.

RABIA TAHIR received the master's degree in computer science from Comsats University Islamabad, in 2017. She is currently pursuing the Ph.D. degree in computer science with Jiangsu University, Zhenjiang.

Her research interests include computer vision, image processing, deep learning, and person re-identification.

MAOZHEN LI received the Ph.D. degree from the Chinese Academy of Sciences. He is currently a Professor with the Department of Electronic and Computer Engineering, Brunel University London. His current research interests include in the areas of high performance computing and big data analytic. He is a member of IET and Fellow of the British Computer Society.