Collective Activity Recognition by Attribute-Based Spatio-Temporal Descriptor

Changhong CHEN\(^{(a)}\), Hehe DOU\(^{(b)}\), Nonmembers, and Zongliang GAN\(^{(c)}\), Member

SUMMARY Collective activity recognition plays an important role in high-level video analysis. Most current feature representations look at contextual information extracted from the behaviour of nearby people. Every person needs to be detected and his pose should be estimated. After extracting the feature, hierarchical graphical models are always employed to model the spatio-temporal patterns of individuals and their interactions, and so can not avoid complex preprocessing and inference operations. To overcome these drawbacks, we present a new feature representation method, called attribute-based spatio-temporal (AST) descriptor. First, two types of information, spatio-temporal (ST) features and attribute features, are exploited. Attribute-based features are manually specified. An attribute classifier is trained to model the relationship between the ST features and attribute-based features, according to which the attribute features are refreshed. Then, the ST features, attribute features and the relationships between the attributes are combined to form the AST descriptor. An objective classifier can be specified on the AST descriptor and the weight parameters of the classifier are used for recognition. Experiments on standard collective activity benchmark sets show the effectiveness of the proposed descriptor.

key words: collective activity recognition, spatio-temporal descriptor, attribute-based spatio-temporal descriptor

1. Introduction

Collective activity recognition in computer vision has received increasing attention in recent years. Collective activity recognition is different from crowd analysis, which tracks person in high-density crowds and aims to search for similar behaviors amongst crowd motion patterns. Collective activities are activities performed by multiple persons, such as queuing in a line, talking together, and waiting at an intersection. It is difficult to be identified by the action of individual person in isolation. People-based methods proposed in recent studies exploit the contextual information of other people nearby the focal person. The spatio-temporal local (STL) descriptor\(^{[1]}\) captures spatial variation over time by head pose and the local surrounding area. They generalize this descriptor in\(^{[2]}\), which significantly increases the flexibility by allowing variable number of regions as well as variable region sizes. The action context (AC) descriptor\(^{[3]}\) is centered on a person (the focal person), and describes the action of the focal person and the behavior of other people nearby. It is suitable when appearance is specific to the target activity. However, an apparent relation descriptor is sensitive to viewpoint change. To solve the problem, the relative action context (RAC) descriptor, proposed in\(^{[4]}\), encodes the “relative” relation. These approaches are people-based descriptors, all of the people should be tracked and their poses need to be estimated. Though the approach of tracking object works well, it is still very hard to automatically detect and track humans in highly cluttered scenes. Moreover, the descriptors differ with the focal person. Therefore, hierarchical graphical models\(^{[3]}\), \(^{[5]}\) are always employed to model the spatio-temporal patterns of individual person and their interactions, and so can not avoid the complex preprocessing and inference operations.

Feature-based methods\(^{[6]}\)–\(^{[8]}\) are more usual for video representation as they detect low-level spatio-temporal features, followed by the definition of the related descriptors. Multiple descriptors can be extracted based on the low-level features, such as extended SURF\(^{[9]}\), and HOG3D\(^{[10]}\). This descriptors are effective for single-person activity representation. However, they can not well describe collective activity due to the semantic information losing. Temporal poselets (TPOS) descriptor proposed in\(^{[11]}\) introduce the feature-based method to collective activity recognition. It analyzes the activation correlation of a bank of (poselet) detectors over time. Though this approach brings a number of advantages such as less computing time, and fewer preprocess steps, the relationship among the people are ignored.

Inspired by the wonderful performance of TPOS descriptor, we propose an attribute-based spatio-temporal (AST) descriptor for collective activity recognition. Attributes represent high-level semantic concepts describing the action. Compared with image-based features, attribute-based descriptor is more suitable for characterizing rich visual temporal-spatial action structures. The framework of our work is shown in Fig.1. Firstly, the spatio-temporal (ST) features, TPOS descriptor and HOG3D, are extracted. Secondly, the attribute of each video block is manually specified. An attribute classifier is trained to model the relationship between the spatio-temporal (ST) features and attribute features. The ST features, attribute features and the relationship between the attributes are connected to form the AST descriptor. An objective classifier can be specified on the AST descriptor and the weight parameters of the classifier.
are used for recognition. For a test video, with the help of the attribute classifier, its AST descriptor can be easily obtained. The predicted label can be obtained by comparing the weight parameters. The AST descriptor retains the advantages of TPOS descriptor and solves the logical information losing by introducing attribute features.

2. ST Feature Extraction

The ST features don’t need to track people and estimate their pose in the video, which avoid complex preprocessing compared with the people-based feature. What’s more, the group detection is dispensable and all the people are concerned in the video. Two kinds of ST features are extracted, TPOS descriptor and HOG3D representation. The TPOS descriptor [11] is extracted as follows:

Step 1: Given an image, poselet detectors [12] are employed as a filter bank. A set of poselets $P$ are detected, which are parts tightly clustered in both appearance and configuration space. $N_p$ kinds of poselets are detected.

Step 2: Each video is divided into $N_w \times N_g \times N_t$ video blocks. Every frame in each block is partitioned in a set of grid cells $g$. The activation of poselet $P$ to the grid cell $g$ is computed as: $v = \frac{\text{area}(g \cap P)}{\text{area}(g \cup P)}$, which is a ratio of the intersection (the white area in Fig. 2) with the union $P$ and $g$. Figure 2 shows the process to obtain the activation value of a poselet (id = 6) to a grid cell. The TPOS descriptor of the video block $k$ is represented as:

$$\text{TPOS}_k = [V_{k,1}, \ldots, V_{k,t}, \ldots, V_{k,T}]$$

where $T$ is the number of video block and $V_{k,t} = [V_{k,t}(p), \ p = 1, 2, \ldots, N_p]$ is the sum activation of the $i^{th}$ frame of the $k^{th}$ video block. $v_{k,t}(p,i)$ is the activation of the $i^{th}$ poselet belong to the $p^{th}$ kind, defined as:

$$V_{k,t}(p) = \sum_{i=1}^{P_p} c_i \ast v_{k,t}(p,i)$$

where $P_p$ is the number of the $p^{th}$ poselet detected in the

$i^{th}$ frame of the video block $k$ and $c_i$ is the score of the $i^{th}$ poselet belong to the $p^{th}$ kind.

The TPOS descriptor embodies the activation correlation of poselet detectors, but does not capture the motion relationship between frames. Therefore, HOG3D [10] is employed to capture the motion information of video block.

3. Collective Activity Recognition by AST Descriptor

Human action described by action attribute are testified to be better than low-level features [13]. The AST descriptor is proposed based on the attribute-based representation.

3.1 Attribute-Based Representation

Attributes in a single person’s action show whether the existence of the environment and state of the people. When more people are in a scene, we should consider not only one person, but also the states of many people. Following the description of [13], we define $M$ different action attributes which encodes the semantic property of the environment factor and the action information. The semantic space is spanned by a basis consist of $M$ attributes $a_i (i = 1, \ldots, M)$. The initial value of $a_i$ is a binary value, either 1 or 0.

Due to the complexity of the collective activities, we manually labeled the attributes for the training dataset. 22 attributes reflecting the motion and scene information are chosen, such as still motion, slow translation, fast translation, cyclic motion, arm free swing, something in hands, take a big step, communication among most people, stand as a line, stand as a column, most people forming a circle, indoor, existing traffic light and occurring at open zone. An example of “queueing” is illustrated in Fig. 3.

3.2 AST Descriptor

The AST descriptor is composed of three parts: the ST features, attribute features and the relationship between the attributes. In order to learn the attribute features automatically for test video, the relationship between attribute-based representation and ST features is necessary. The spatial information can be reflected by the the relationship between the attributes. The AST descriptor can be obtained by the following steps:

Step 1: The relationship between attribute-based representation and ST features can be easily built by attribute classifiers. Suppose $x \in \mathcal{X}^d$ is the ST feature and $a \in \mathcal{A}^M$ the attribute-based representation. $M$ attribute classifiers can be built to realize:
\begin{equation}
X^d \rightarrow A^M
\end{equation}

Each classifier maps \(x\) to the \(j^\text{th}\) attribute of \(A^M\). Each attribute corresponds to an attribute classifier. For each classifier, existence of attribute \(a_i = 1\) is treated as the positive examples and absence of attribute \(a_i = 0\) as the negative examples. After the attribute classifiers established, their decision values can be used to refresh the initialized attribute features. The decision values of the attribute classifiers is more credible.

Step 2: The task is to find the optimal label \(y\) for the given video. By considering the ST features and attribute features, the proposed approach is formulated as:

\begin{equation}
f_\omega(x, y) = \omega_1 \varphi_1(x; y) + \sum_j \omega_{aj}^T \varphi_2(x, a_j; y) + \sum_{jk} \omega_{aj, ak}^T \varphi_3(a_j, a_k; y)
\end{equation}

where \(\varphi_1(x; y)\) corresponding to the ST features, \(\varphi_2(x, a_j; y)\) is the attribute feature, which corresponds to the decision values of the attribute classifiers, \(\varphi_3(a_j, a_k; y)\) represents the dependencies between the \(j^\text{th}\) attribute and the \(k^\text{th}\) attribute. \(\varphi_3(a_j, a_k; y)\) is defined as four-dimensional vector which indicates one of four possible configurations (0, 0), (0, 1), (1, 0), (1, 1) of \((a_j, a_k)\). For instance, if the video has both the attribute \(a_i\) and \(a_{i+1}\), \(\varphi_3(a_i, a_{i+1}; y) = (0, 0, 0, 1)\). In the procedure of model learning, the goal is to estimate the weight parameters of appearance model \(\omega_a\), intra-class attribute model \(\omega_{ai}\), and the inter-class attribute model \(\omega_{ai, ak}\). The above equation can be rewritten as follows:

\begin{equation}
f_\omega(x, y) = w^T \Phi(x, y, a)
\end{equation}

where \(w\) is the model parameter and \(\Phi(x, y, a)\) is the AST descriptor, defined as:

\begin{equation}
w = \begin{bmatrix} w_s \\ w_{aj} \\ w_{ak} \\ w_{ai, ak} \end{bmatrix}, \quad \Phi(x, y, a) = \begin{bmatrix} \varphi_1(x; y) \\ \varphi_2(x, a_j; y) \\ \varphi_3(a_j, a_k; y) \end{bmatrix}
\end{equation}

The three parts of the AST descriptor reflect the motion and the semantic information of the collective activities. Although the attributes should be manually labeled for the training video, the attributes of the test video is easy to be estimated according to the attribute classifiers.

### 3.3 Model Learning

The training set is divided into two subsets, where all of the training instances in class \(m\) are considered as a positive class, and the rest as a negative class (similar to the one-vs-rest setting). To learn an optimal weight parameter \(\omega\), the objective function can be converted to a regularized learning problem as follows:

\begin{equation}
\arg\min_{\omega} f_\omega(x, y) = \sum_m (0, 1 - y_m f_\omega(x_m))
\end{equation}

where the second term is a soft-margin, and we can obtain the minimum of the objective function to get the parameter \(\omega\). Stochastic gradient descent method or coordinate descent method, SVM [14] can be utilized to solve the problem.

With the model parameter \(\omega\), the inference procedure is to find the best label set \(y^*\) for a test video \(x\). It aims to solve the optimization problem as:

\begin{equation}
y^* = \arg\max_y f_\omega(x, y)
\end{equation}

### 4. Experiments and Discussion

In order to evaluate the performance of the proposed AST descriptor, we conduct experiments on the collective activity dataset (CAD1) [1] and a newly collective activity dataset (CAD2) [2]. We randomly select half of the video clips to form the training set, and the rest are used for testing. Forty experiments are executed by changing the training and test samples.

The comparison average results of the AST descriptor and other collective activity descriptors are shown in Table 3. The confusion matrices of the two datasets are displayed in Fig. 4. The major error occurs in action “waiting”, whose recognition rates are 50% for CAD1 and 55% for CAD2. Generally, “waiting” is similar with “crossing”, which is composed of “waiting” and “crossing”. Another two confusables are “walking” and “crossing” for CAD1, because “crossing” is a kind of “walking” in essence. The actions in CAD2 are easier to be distinguished because more vigorous actions are contained.

### 5. Conclusion

This paper proposes a new representation for collective ac-
Table 1 Comparison results of CAD1 and CAD2.

|       | ST   | AST  | TPOS [11] | STL [1] | RSTV [2] | AC [3] | AC+full model [5] |
|-------|------|------|-----------|--------|----------|--------|-------------------|
| CAD1  | 57.4 | 73.4 | -         | 65.9   | 67.2     | 68.2   | 72.0              |
| CAD2  | 62.8 | 81.7 | 72.9      | -      | 71.7     | -      | 85.8              |

Fig. 4 Confusion matrices of the AST descriptor.

Activity, the AST descriptor, which keep the advantages of the TPOS descriptor and offset the loss of logical information occasioned by introducing attribute features. The AST descriptor needs no complex preprocessing operations, such as tracking and pose evaluation, and recognises the type of action in a video sequence. The classifier of the AST descriptor is simple, while person-based representation needs complex model to obtain comparable results. The AST descriptor is easy to realize and gets encouraging performance. Future work will be focused on automatically learning the attributes.

Acknowledgements

This work is supported by the NSF of China Nos. 61172118 and 61471201, Jiangsu Province Universities Natural Science Research Key Grant Project No.13KJA510004, Open Project of State Key Laboratory of Novel Software Technology (Nanjing University) No. KFKT2014B10, Natural Science Foundation of Jiangsu Province BK20130867.

References

[1] W. Choi, K. Shahid, and S. Savarese, “What are they doing?: Collective activity classification using spatio-temporal relationship among people,” International Workshop on Visual Surveillance, IEEE, 2009.

[2] W. Choi, K. Shahid, and S. Savarese, “Learning context for collective activity recognition,” IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp.3273–3280, 2011.

[3] T. Lan, Y. Wang, G. Mori, and S.N. Robinovitch, “Retrieving actions in group contexts,” International Workshop on Sign Gesture Activity, QIL, vol.6553, pp.181–194, 2012.

[4] T. Kaneko, M. Shimosaka, S. Odashima, R. Fukui, and T. Sato, “Viewpoint invariant collective activity recognition with relative action context,” European Conference on Computer Vision, vol.7585, pp.253–262, Springer, 2012.

[5] S. Khamis, V.I. Morariu, and L.S. Davis, “Combining per-frame and per-track cues for multi-person action recognition,” European Conference on Computer Vision, vol.7572, pp.116–129, 2012.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” International workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.65–72, IEEE, 2005.

[7] I. Laptev, “On space-time interest points,” International Journal of Computer Vision, vol.64, no.2-3, pp.107–123, 2005.

[8] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time interest points,” IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp.1948–1955, 2009.

[9] G. Willems, T. Tuytelaars, and L. Van Gool, “An efficient dense and scale-invariant spatio-temporal interest point detector,” European Conference on Computer Vision, vol.5303, pp.650–663, 2008.

[10] L. Ballan, M. Bertini, A.D. Bimbo, L. Seidenari, and G. Serra, “Effective codebooks for human action representation and classification in unconstrained videos,” IEEE Transactions on Multimedia, vol.14, no.4, pp.1234–1245, 2013.

[11] M. Nabi, A. Del Bue, and V. Murino, “Temporal poselets for collective activity detection and recognition,” ICCV Workshops C Understanding Human Activities Context and Interactions, pp.500–507, 2013.

[12] L. Bourdev and J. Malik, “Poselets: Body part detectors trained using 3d human pose annotations,” IEEE International Conference on Computer Vision, pp.1365–1372, 2009.

[13] J. Liu, B. Kuipers, and S. Savarese, “Recognizing human actions by attributes,” IEEE International Conference on Computer Vision and Pattern Recognition, pp.3337–3344, 2011.

[14] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol.2, no.3, pp.107–123, 2002.