Human activity recognition using overlapping multi-feature descriptor

S.Y. Cho and H.R. Byun

An efficient overlapping multi-feature descriptor and classification scheme for human activity recognition is introduced. The descriptor is constructed by overlapping global feature combinations of multi-frames using a Hankel matrix representation. The descriptor captures the local and temporal information while overcoming the limitations of global features using an overlapping combination scheme. In addition, a random forests classifier is used to cope with noise in the descriptor that can be obtained from no-activity frames in a video. Using this framework, it is shown that the approach outperforms the state-of-the-art methods using the KTH dataset and a much more complex human interaction dataset.

Introduction: Various features have been developed to recognise human activity in videos. Early approaches to recognising activities were based on silhouette-based tracking or a motion shape template. However, these global features cannot handle real-world videos that contain complex and changing backgrounds and large variations in appearance and motion. To overcome these limitations, the most recent work has used local features to model activities. Laptev [1] introduced space-time interest points similar to the Harris operator and applied this feature to recognise walking. Since then, space-time interest point-based approaches [2–4] have become widely used for activity recognition. These approaches successfully represent the activity as a bag of space-time features. Niebles *et al.* [2] presented an unsupervised learning method using a bag of video words model that was constructed using a probabilistic Latent Semantic Analysis (pLSA) model. Schechtman and Irani [3] introduced a behaviour-based similarity measure between two different space-time intensity patterns in activity video segments. Jhuang *et al.* [4] applied a biological model of motion processing based on space-time feature detectors. However, these are unable to capture smooth activities because of a lack of temporal information, and potentially valuable global information is disregarded. Furthermore, they still have limitations on real-world video that can be represented by a noisy descriptor owing to complex backgrounds and much variation in actors or camera motion.

In this Letter, we propose an efficient overlapping multi-feature descriptor for activity recognition. The activity video is represented as a proposed descriptor obtained from the activity matrix that is built by overlapping global feature combinations using a Hankel matrix representation. Therefore, our descriptor can capture local and temporal information while overcoming the limitations of global features through overlapping multiple global features. In addition, by using a random forests classifier with our descriptor, we can deal with the noisy data that originates from labelling of the no-activity region. We show that our method can improve recognition results in both simple and complex cases by comparing the results obtained from our method with those obtained using state-of-the-art methods.

Descriptor construction: To construct the proposed descriptor, we first extract the histograms of oriented gradient (HoG) and histograms of optical flow (HoF) from all frames in a video clip. HoG is built by discretising the gradients into five bins: horizontal, vertical, two diagonal orientations and no-gradient. HoF is also built by discretising the optical flows into five bins: left, right, up, down and no-motion. The two histograms are normalised and the initial descriptor h is generated by combining two histograms. Given $h_i^0, h_i^1, \ldots, h_i^{n_i-1}$ extracted from n_i frames in each video i, we represent these descriptors as matrix H_i that is similar to a Hankel matrix:

$$H_i = \begin{bmatrix} h_i^0 & h_i^1 & h_i^2 & \cdots & h_i^{n_i-r_i+1} \\ h_i^1 & h_i^2 & h_i^3 & \cdots & h_i^{n_i-r_i+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ h_i^{r_i-1} & h_i^{r_i} & h_i^{r_i+1} & \cdots & h_i^{n_i} \end{bmatrix} \quad (1)$$

where the number of columns of H_i is the number of frames n_i, and the number of rows is the r_i. r_i determines the degree of overlap among frame descriptors. We choose $r_i = 4$ to avoid excessive overlap. Note that the size of the matrix H_i for each video i is manifold. Although an original Hankel matrix is a square matrix, it is difficult to construct the H_i as a square matrix because each video usually contains a different number of frames. To obtain the descriptor of equal size from H_i, we use the orthogonal basis obtained from the singular value decomposition (SVD) of $H_iH_i^T$:

$$[U, S, U^T] = svd(H_iH_i^T) \quad (2)$$

where S is a diagonal matrix and U is an orthogonal matrix. The columns of U are orthonormal eigenvectors associated with the eigenvalues of S. The final descriptor d_i for video i is built by projecting the $H_iH_i^T$ on the first eigenvector of U:

$$d_i = u_l^T H_i H_i^T \quad (3)$$

Human activity recognition: In our approach, an activity is represented by the proposed overlapping multi-feature descriptor d. For activity recognition, we are given a training dataset with descriptors of N_D video clips, $D = \{d_1, d_2, \ldots, d_{N_D}\}$. We use a random forests classifier to recognise human activities. A random forest is an ensemble classifier that consists of many decision trees grown using some form of randomisation. The benefits of a random forests classifier are numerous, in particular it runs efficiently on datasets with a considerable amount of labelling noise. In the training procedure, the random forest with N_T trees is built by choosing a random subset D' from the descriptors of training data D. Each tree is fully grown by calculating the best split for each node of the tree. In the testing procedure, an activity descriptor of a test video is passed down all random trees, and a test video d is classified using (4). In (4), $P_t(f(d) = c|d)$ is an *a posteriori* probability at the leaf node of each tree t for each activity class $c \in C$ and $f(d)$ is the class-label for video d:

$$f^*(d) = \arg\max_c \frac{1}{N_T} \sum_{t=1}^{N_T} P_t(f(d) = c|d) \quad (4)$$

Experimental results: To construct the random forest, we set the number of trees in the forest to $N_T = 36$ and the maximum tree depth to 10 because this setting has been demonstrated to show the best average performance. We first evaluate the performance of our method and compare our results to those obtained using state-of-the-art methods [5, 6] on the KTH dataset [7]. The KTH dataset consists of six human actions: walking, jogging, running, boxing, hand waving and hand clapping. The sequence for each action is performed several times by 25 subjects in four different scenarios: outdoor, outdoors with scale variation, outdoors with different clothes and indoors. The overall average precision of the proposed approach is 94.55%. Table 1 shows that our method outperforms the state-of-the-art methods using the standard settings [7].

Table 1: Comparison of overall average precision of our method and state-of-the-art methods for KTH dataset

Method	Average precision
Schuldt *et al.* [7]	71.5
Niebles *et al.* [5]	91.3
Wang *et al.* [6]	94.2
Our method	**94.55**

Table 2: Comparison of average precision per interaction class for TV human interaction dataset

Method	Hand shake	High five	Hug	Kiss	Average
ID [7]	**0.48**	0.32	0.42	0.32	0.38
SL [7]	0.44	0.33	0.44	0.36	0.39
Our method	0.36	**0.52**	**0.56**	**0.68**	**0.53**

ID is an independent classification that uses the information of one of the people performing the interaction. SL is a structured classification that applies a structured learning framework using information from all the people involved in the interaction

Our method was also tested using a more challenging TV human interaction dataset [8]. The dataset contains four human interactions: hand shake, high five, hug and kiss. Because the video clips were compiled from 23 different TV shows, there are large variations within video clips of the same class. In other words, each scene has various backgrounds, number of actors, scales, camera angles and abrupt viewpoint changes. We compared the average precision results per interaction class

with Patron-Perez *et al.* [8]. In [8], they use the feature descriptor obtained from the upper body and head orientation of a person. Table 2 shows a performance comparison of our method with [8]. Our method outperforms the performance of [8], with an improvement of 15 and 14% over the independent (ID) and structured learning (SL) classification results, respectively.

Fig. 1 shows the true and false positives for each interaction obtained using our method. As mentioned above, the dataset includes variations in the number of actors, backgrounds, and camera views, as shown in Fig. 1. We observed that hand shakes tend to be recognised as high fives and vice versa. This could be because both of the two interactions involve the holding out of hands and it is difficult to classify and differentiate between the two interactions under such limited variation. Furthermore, in many cases, the hand shakes or high fives tend to be recognised as hugs or kisses when there is extensive overlap of bodies and faces between the interacting people.

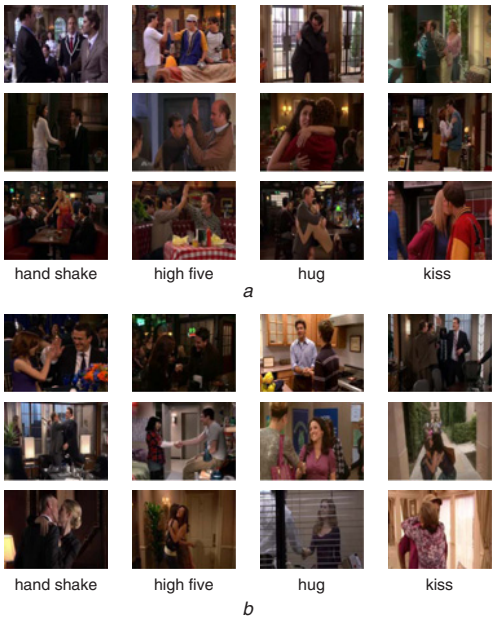

Fig. 1 *True and false positives for each interaction class obtained using our method*

a True positives
b False positives

Conclusion: We propose an efficient framework for human activity recognition that uses an overlapping multi-feature descriptor and a random forests classifier. Our descriptor captures local and temporal information using overlapping global feature combinations. The random forests classifier is used with the descriptor to handle the noisy portions of the descriptor. We have demonstrated that our proposed method outperforms the state-of-the-art methods for both the KTH dataset and the much more challenging TV human interaction dataset. In future work, we will exploit a more robust feature descriptor using various types of feature combinations on a challenging real-world activity dataset.

Acknowledgments: This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Centre) support programme supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-(C1090-1121-0008)).

12 August 2011
doi: 10.1049/el.2011.2550
One or more of the Figures in this Letter are available in colour online.

S.Y. Cho and H.R. Byun (*Department of Computer Science, Yonsei University, ShinChon-Dong Sudaemoon-Ku, Seoul 120-749, Republic of Korea*)

E-mail: sycho22@yonsei.ac.kr

References

1 Laptev, I.: 'On space-time interest points', *Int. J. Comput. Vis.*, 2005, **64**, (2/3), pp. 107–123
2 Niebles, J., Wang, H., and Fei-Fei, L.: 'Unsupervised learning of human action categories using spatial-temporal words', *Int. J. Comput. Vis.*, 2008, **79**, (3), pp. 299–318
3 Shechtman, E., and Irani, M.: 'Space-time behavior based correlation'. Proc. Int. Conf. Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, Vol. 1, pp. 405–412
4 Jhuang, H., Serre, T., Wolf, L., and Poggio, T.: 'A biologically inspired system for action recognition'. Int. Conf. on Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8
5 Niebles, J.C., Chen, C.-W., and Fei-Fei, L.: 'Modeling temporal structure of decomposable motion segments for activity classification'. European Conf. on Computer Vision, Hersonissos, Greece, 2010, pp. 1–14
6 Wang, H., Klaser, A., Schmid, C., and Liu, C.-L.: 'Action recognition by dense trajectories'. Proc. Int. Conf. Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 2011, pp. 3169–3176
7 Schuldt, C., Laptev, I., and Caputo, B.: 'Recognizing human actions: a local svm approach'. Proc. Int. Conf. on Pattern Recognition, Cambridge, UK, 2004, Vol. 3, pp. 32–36
8 Patron-Perez, A., Marszalek, M., Zisserman, A., and Reid, I.: 'High five: recognising human interactions in TV shows'. Proc. British Machine Vision Conf., Aberystwyth, UK, 2010, pp. 1–11