Human-action recognition module for the new generation of augmented reality applications

Ana I. Maqueda, Carlos R. del-Blanco, Fernando Jaureguizar, and Narciso García

***Abstract*— Augmented reality is becoming more and more popular due to the countless number of practical applications. A key element is the understanding of the scene and the involved human activities to be able to offer a rich interaction with the world via virtual actions and elements. For this purpose, a new vision-based human-action recognition module has been developed to be integrated with the new generation of augmented reality devices, which is based on Spatio-Temporal Interest Points and the Citation K-Nearest Neighbor classifier.**

Keywords—augmented reality, human action recognition, spatio-temporal features, STIP, C-KNN.

I. INTRODUCTION

Augmented reality is an emerging technology with a countless number of practical applications. The last advances in machine learning, computer vision, and smart electronic devices are making possible its integration in everyday life. The goal of augmented reality is to fuse information from the real world with synthetic and virtual information to experience a new level of interaction among the reality, the knowledge of the humankind, and us (the human beings). An essential element for this purpose is to sense and understand the world [1], along with the human activities that take place in it [2]. Vision-based human action recognition techniques are important tools to achieve it.

A wide range of methods focused on the visual recognition of human actions have been developed. In particular, techniques based on local spatio-temporal features are one of the most popular to characterize human actions in video sequences [3]. In order to extract such features, Spatio-Temporal Interest Points (STIP) are detected and robustly characterized by an image descriptor. Recently, many works have proposed to track STIP features to create new descriptors from their trajectories [4], using even a combination of several descriptors to characterized each STIP [5]. In order to efficiently combine in a compact representation multiple STIPS that are spread temporally and spatially, a Bag of Words (BoW) approach is commonly used. The resulting feature vector is then delivered to a matching algorithm or machine learning technique, such as K-Nearest Neighbor (K-NN) or Support Vector Machine (SVM) to perform the action recognition.

In this paper, a new human-action recognition module that achieves a better recognition accuracy regarding other state-of-the-art approaches is presented. The human activities in a video flow are represented by local spatio-temporal features, created by detecting STIPs that are later characterized by a combination of HOG-HOF (Histograms of Oriented Gradients-Histograms of Optical Flow) descriptors, and finally combined using a BoW approach [6]. The resulting feature descriptors are used as inputs for a classifier, called Citation K-Nearest Neighbor (C-KNN) [7], which is especially robust for the BoW-based human activity representation, where a high number of outliers/misleading data can be present.

Fig. 1. Proposed human-action recognition module.

II. HUMAN-ACTION RECOGNITION MODULE

The proposed human-action recognition module is outlined in Fig. 1.

A. *Spatio-Temporal Interest Points (STIP)*

In order to extract local spatio-temporal features, first STIP are detected by using 3D Harris detector. Then, the neighborhood of each STIP is described by computing HOG and HOF features, which represent appearance and motion information, respectively [6]. An example of the STIP detection is shown in Fig. 2.

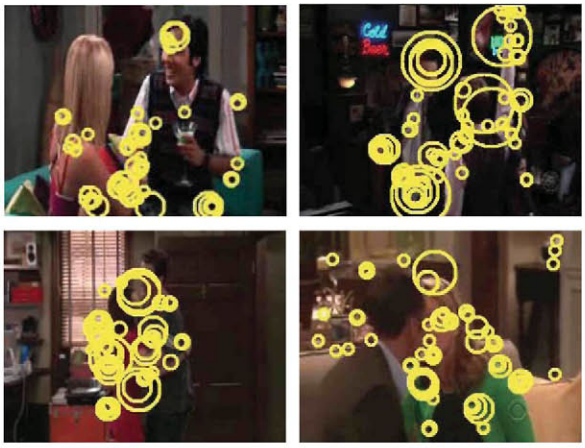

Fig. 2. Detected STIP for four video frames corresponding to the human actions: *hand shake* (upper-left), *high five* (upper-right), *hug* (bottom-left), and *kiss* (bottom-left).

B. Bag of Features (BoW)

After the STIP detection and description, a video sequence is represented by a set of feature vectors. To create a compact and vectorized representation, a BoW approach is used. To that end, a training set of features is clustered by using the K-means algorithm, where the resulting cluster centroids represent a visual dictionary. Then, features from new video sequences are assigned to the nearest visual word using the Euclidean distance. Finally, a histogram of visual words is computed.

C. Classification through Citation K-NN

Citation-KNN algorithm was introduced to adapt K-Nearest Neighbor (K-NN) to Multiple-Instance Learning (MIL) problems [7], where the BoW strategy represents an example of these. In this context, a classifier learns from a training set of *bags* that contains multiple feature vectors. Each bag has an associated label, but the labels of the individual instances are unknown.

C-KNN is based on two main ideas. The first one consists in defining a new distance function between bags, called minimal Hausdorff distance [7]. Given two bags of instances $A=\{a_1,...,a_n\}$ and $B=\{b_1,...,b_m\}$, the mathematical expression is defined as:

$$h_1(A,B) = \min_{a_i \in A} \min_{b_j \in B} \left\| a_i - b_j \right\|. \quad (1)$$

The second idea consists in predicting the label of a given bag b by considering, not only the bags contained in its nearest neighborhood (*references*), but also the bags that contain b in their nearest neighborhood (*citers*). Once the R-nearest references and the C-nearest citers are computed, if the number of positive bags are larger than the number of negative bags, the label of the bag b is predicted as positive, and otherwise negative.

To deal with the ongoing multi-class problem of human action recognition, one-against-one strategy has been adopted. This way, a voting strategy is applied to the predictions extracted from binary classifiers to make a final decision for every video sequence.

III. Results

In this section, the proposed learning approach is evaluated on the TVHI dataset [8]. This dataset consists of four types of human actions: *handshake*, *highfive*, *hug* and *kiss*. Each class contains 50 video sequences, which are divided into two sub-sets: a training set containing the 80% of the sequences, and a test set containing the 20% of the sequences.

Table I shows a comparison with other state-of-the-art approaches, by using the *Average accuracy* metric, which is defined as follows:

$$\text{Average accuracy (\%)} = \frac{\text{Total number of correct actions}}{\text{Total number of actions}} \times 100. \quad (2)$$

The first method [9] combines STIP and HOG-HOF descriptors with BoW, and once the visual words have been computed, they are filtered to capture the most meaningful data. In this case, an SVM classifier is used. The second approach [4] is based on dense trajectories and Motion Boundary Histograms (MBH). A BoW together with a SVM classifier is used. Finally, K-NN is used to be compared with C-KNN. It can be observed that our system outperforms the others, and only the system described in [4] is relatively close.

TABLE I. AVERAGE ACCURACY OBTAINED WITH DIFFERENT LEARNING FRAMEWORKS.

Method	Accuracy
STIP + HOG-HOF + filtered BoW + SVM [9]	50.5%
Dense trajectories + MBH + BoW + SVM [4]	56.0%
STIP + HOG-HOF + BoW + K-NN	25.0%
STIP + HOGHOF + BoW + C-KNN	**57.5 %**

IV. Conclusion

A new human-action recognition module is presented in this paper. It is based on local spatio-temporal features, which are extracted from video sequences by first detecting STIP, and computing HOG-HOF descriptors. For the classification, a BoW approach is adopted together with the Citation K-NN classifier. This classification methodology outperforms other state-of-the-art approaches, proving its suitability for human action recognition oriented to augmented reality.

References

[1] D. Geronimo, and H. Kjellstrom, "Unsupervised Surveillance Video Retrieval Based on Human Action and Appearance," Internation Conference on Pattern Recognition, pp.4630,4635, 24-28, August 2014.

[2] K.T. Song, and W.J. Chen, "Human activity recognition using a mobile camera," International Conference on Ubiquitous Robots and Ambient Intelligence, pp.3,8, 23-26, November 2011.

[3] I. Everts, J.C. van Gemert, and T. Gevers, "Evaluation of Color Spatio-Temporal Interest Points for Human Action Recognition," IEEE Transactions of Image Processing, vol.23, no.4, pp.1569,1580, April 2014.

[4] W. Heng, A. Klaser, C. Schmid, and L. Cheng-Lin, "Action recognition by dense trajectories," IEEE Conference on Computer Vision and Pattern Recognition, pp.3169,3176,20-25, June 2011.

[5] C. Li, B. Su, Y. Liu, H. Wang; J. Wang, "Human action recognition using spatio-temoporal descriptor," International Congress on Image and Signal Processing , vol.1, pp.107,111,16-18, December 2013.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human action from movies," IEEE Conference on Pattern Recognition, pp.1,8,23-28, June 2008.

[7] J. Wang, and J.D. Zucker, "Solving multiple-learning-instance problem: a lazy learning approach," International Conference on Machine Learning, pp.1119-1125, June 2000.

[8] A. Patron-Perez, M. Marskalez, I. Reid, and A. Zisserman, "Structured Learning of Human Interactions in TV shows," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.34, no.12, pp.2441,2453, December 2012.

[9] B. Zhang, F.G.B. De Natale, and N. Conci, "Recognition of social interactions based on feature selection from visual codebooks," IEEE International Conference on Image Processing, pp.3557,3561, 15-18, September 2013.