BoVDW: Bag-of-Visual-and-Depth-Words for Gesture Recognition

Abstract

We present a Bag-of-Visual-and-Depth-Words (BoVDW) model for gesture recognition, an extension of the Bag-of-Visual-Words (BoVW) model that benefits from the multimodal fusion of visual and depth features. In our BoVDW model, state-of-the-art RGB and Depth features are analysed and combined in a late-fusion fashion. The method is integrated in a continuous gesture recognition pipeline, where Dynamic Time Warping (DTW) algorithm is used to perform prior segmentation of gestures. Results of the method in public data sets, within our gesture recognition pipeline, show better performance of the proposed BoVDW with late-fusion in comparison to early-fusion and standard BoVW model.

1. Introduction

Nowadays, BoVW is one of the most used approaches in Computer Vision, commonly applied in image retrieval or image classification scenarios. This methodology is an evolution of *Bag-of-Words* (BoW) [6], a method used in document analysis, where each document is represented using the apparition frequency of each word in a dictionary. In the image domain, the words become *visual words* and are elements of a certain *visual vocabulary*. First, each image is decomposed into a large set of patches, either using some type of spatial sampling (grids, sliding window, segmentation, etc.) or detecting points with interesting properties (corners, salient regions, maximally stable extreme regions, etc.). Each patch is then described, obtaining a numeric *descriptor*. A set of N representative *visual words* are selected by means of a clustering process over the *descriptors*, where N is the cardinality of the *visual vocabulary*. Once the *visual vocabulary* is defined, each image can be represented by a global histogram containing the frequencies of *visual words*. Finally, this histogram can be used as input for any classification technique (i.e. $k-$Nearest Neighbor or SVM) [3, 7]. In addition, extensions of BoW from still images to image sequences have been recently proposed, defining Spatio-Temporal-Visual-Words (STVW) (i.e. in the context of human action recognition) [8].

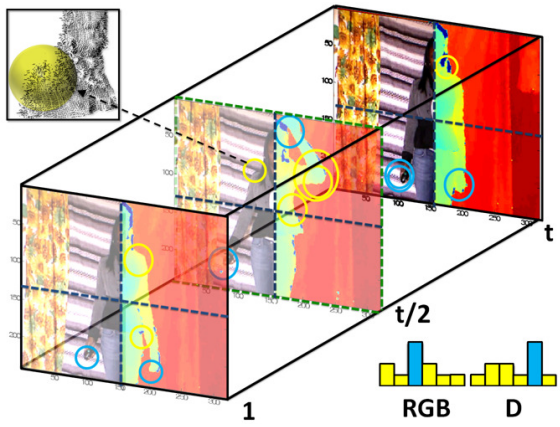

Figure 1. Conceptual scheme of the BoDVW approach. Interest points in RGB and depth images are depicted as circles. Color of the circles indicates the assignment to a visual word in the shown histogram –computed over one spatio-temporal bin. Spatio-temporal bins limits are represented by dashed lines in blue and green. respectively). A detailed view of the normals of the depth image is shown in the upper-left corner.

Since its release in late 2010, the Microsoft Kinect sensor caused a frenetic expansion in the computer vision field. Kinect is a low cost sensor which is able to capture depth information of the scene, in addition to the RGB image provided by a camera, providing what is named RGB-D images (RGB plus Depth). This depth information has been particulary exploited for human body segmentation and tracking. Girshick and Shotton [11] presented one of the greatest advances in the extraction of the human body pose using RGBD, which is provided as part of the Kinect human recognition framework. Moreover, motivated by the information provided by depth maps, several 3-D descriptors have

been recently developed [9, 10], which are based on codifying the distribution of normal vectors among regions in the 3D space.

In this paper, we present a Bag-of-Visual-and-Depth-Words (BoVDW) approach, which is an extension of the BoVW approach that takes profit of multimodal RGB-D images by combining information of both RGB images and depth maps. We also propose a new depth descriptor which takes into account the distribution of normal vectors respect the camera position, as well as the rotation respect the roll axis of the camera. In order to do a complete evaluation of the presented approach, we compare the performances achieved with state-of-the-art RGB and depth features separately, and combining them in a late-fusion fashion. All experiments are run in the proposed framework using the public data set provided by the ChaLearn Gesture Challenge[1] in the context of gesture recognition. Finally, the presented BoVDW approach is integrated in a fully-automatic system for gesture recognition, which uses DTW for the prior segmentation of gestures in a sequence. Results of the method show better performance of the proposed BoVDW with late-fusion in comparison to early-fusion and standard BoVW model. Summarizing, the contributions of our work are: a) The multi-modal BoVDW model, b) A new depth descriptor, c) the performance analysis of state-of-the-art RGB and depth descriptors within the BoVDW model, d) the analysis of late-fusion in comparison to early-fusion in the BoVDW model, and e) the design of a fully automatic system for continuous gesture recognition using the proposed BoVDW model applied to public data sets.

The rest of the paper is organized as follows: Section 2 presents the BoVDW method for gesture recognition. Section 3 compares our proposal with state-of-the-art approaches on the ChaLearn data data set. Finally, Section 4 concludes the paper.

2 BoVDW

In this section, we present the Bag-of-Visual-and-Depth-Words approach for Gesture Recognition. The BoVDW pipeline is shown in Figure 2 (blue pipeline), and Figure 1 contains a conceptual scheme of the approach. The steps of the procedure are described next: keypoint detection, keypoint description, histogram computation, and classification. Finally, we present the application of the BoVWD approach to a Gesture Recognition system (green pipeline in Figure 2).

[1] http://gesture.chalearn.org/

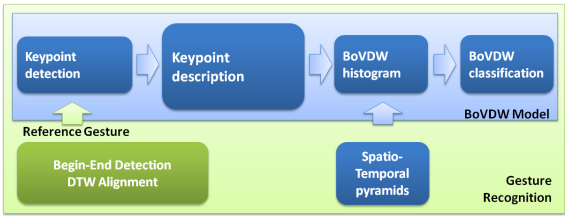

Figure 2. BoVDW-based Gesture Recognition.

2.1 Keypoint detection

The first step of BoW-based models consists of selecting a set of points in the image/video with special properties. Since a dense sampling over the whole video could result in a huge amount of points, keypoint detection is mainly used in order to reduce the computational complexity of this step. In our case, we use the Spatio-Temporal Interest Point (STIP) detector [4], which is an extension of the well-known Harris detector in the temporal dimension. The STIP detector first computes the second-moment 3×3 matrix μ of first order spatial and temporal derivatives. Then, the detector looks for regions in the image with significant eigenvalues λ_1,λ_2,λ_3 of μ, combining the determinant and the trace of μ:

$$H = |\mu| - k \cdot T_r(\mu)^3, \quad (1)$$

where $|.|$ corresponds to the determinant, $T_r(.)$ computes the trace, and k stands for a relative importance constant factor. As we have multimodal RGB-D data, we apply the STIP detector separately on the RGB and Depth volumes, so we get two sets of interest points P_{RGB} and P_D.

2.2 Keypoint description

At this step, we want to describe the interest points detected in the previous step. On one hand, for P_{RGB} we compute state-of-the-art RGB descriptors, including HOG, HOF, and their concatenation HOG/HOF [5]. On the other hand, for P_D we test the VFH descriptor and propose the VFHCRH, detailed below.

2.2.1 VFHCRH

The recently proposed PFH and FPFH descriptors [1] describe each point of the cloud with a histogram describing the mean curvature around it. Both PFH and FPFH provide $n \cdot 6D$ pose invariant histograms, being n the number of points in the cloud. Following their principles, VFH describes each object with one descriptor of 308 bins, variant to object rotation around pitch and

yaw axis. However, VFH is invariant to rotation about the roll axis of the camera.

In contrast, CVFH describes each object using a different number of descriptors r, where r is the number of stable regions found on the object. Each stable region is described using a non-normalized VFH histogram and a Camera's Roll Histogram (CRH), and the final object description includes all region descriptors. CRH is computed by projecting the normals at each point onto a plane that is orthogonal to the vector between the centroid of the region and the camera center, under orthographic projection. Then, the histogram is computed by counting the projected angles relatives to the vertical vector of the camera plane.

In order to avoid descriptors of arbitrary lengths for different objects, we describe the whole object or image region using VFH. In addition, a 92 bins CRH is computed for encoding $6DOF$ information. The concatenation of both histograms results in the proposed VFHCRH descriptor of 400 bins shown in Figure 3.

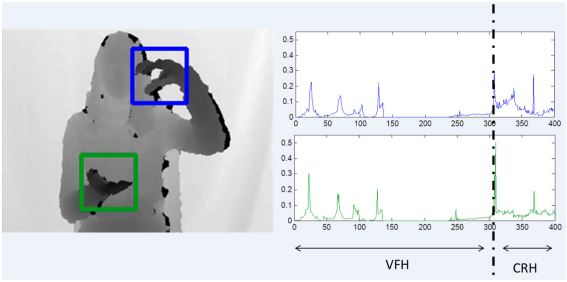

Figure 3. VFHCRH descriptor: Concatenation of VFH and CRH histograms resulting in 400 total bins.

2.3 BoVDW histogram

Once we have described all the detected points, we build our vocabulary of N visual/depth words by applying a clustering method over all the descriptors. Hence, the clustering –k-means in our case– defines the words from which a query video will be represented, shaped like a histogram h that counts the frequencies of each word.

Additionally, in order to introduce geometrical and temporal information, we apply spatio-temporal pyramids. Basically, spatio-temporal pyramids consist of dividing the video volume in b_x, b_y, and b_t bins along the x, y, and t dimensions of the volume, respectively. Then, $b_x \times b_y \times b_t$ separate histograms are computed with the points lying in each one of these bins, and concatenated jointly with the general histogram computed by using all points.

These histograms define the model for a certain class of the problem –in our case, a certain gesture. Since we deal with multimodal data, we build different vocabularies for the RGB-based descriptors and the Depth-based ones, and obtain then a different histogram for each kind of features, h^{RGB} and h^{D}. Therefore, the information given by the different features is merged in the next and final classification step, hence using *late fusion*.

2.4 BoVDW-based classification

The final step of the BoVDW approach consists of predicting the class of the query video. For that, any kind of multi-class supervised learning technique can be used. In our case, we use a simple k-Nearest Neighbor classification, computing the complementary of the histogram intersection as distance:

$$d^F = 1 - \sum_i \min(h^F_{model}(i), h^F_{query}(i)). \quad (2)$$

Finally, in order to merge the histograms h^{RGB} and h^{D}, we compute the distances d^{RGB} and d^{D} separately, and compute a weighted sum: $d = (1-\alpha)d^{RGB} + \alpha d^{D}$.

2.5 Gesture Recognition system

In order to test the BoVDW model representation, we designed a continuous gesture recognition system. First, *Dynamic Time Warping* is used to detect a gesture of reference which splits the multiple gestures to be recognized. Then, each segmented gesture is classified using the BoVDW pipeline described before. These steps are also illustrated in the green pipeline shown in Figure 2.

3 Experiments

In order to present the results, first, we discuss the data, methods, and evaluation metrics of the comparative.

3.1 Data

We used the ChaLearn [2] development data set provided from the CVPR2011 Workshop on Gesture Recognition. It consists of video sequences captured with the Kinect device, providing both RGB and depth images. The sequences are organized in 20 batches, each one of them including 100 recorded gestures grouped in sequences containing from 1 to 5 gestures,

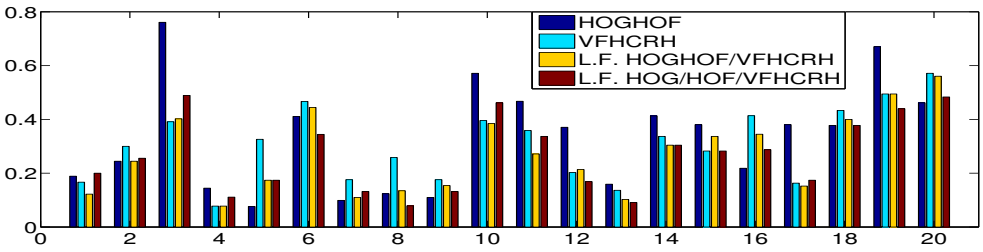

Figure 4. Bar plot showing the performance of the best RGB and depth descriptors separately, as well as the 2-fold and 3-fold late fusion of them. X axis represent different batches, and Y axis represents the MLD of each batch.

performed by the same user. The gestures are drawn from a small vocabulary of 8 to 15 unique gestures, which we call a *lexicon*[2]. Since the aim of this challenge is to develop a one-shot-learning gesture recognition system, just one training sample per gesture is provided.

3.2 Methods

For the experiments shown in this section, the vocabulary size was set to $N = 200$ words for both RGB and depth cases. For the spatio-temporal pyramids, the volume was divided in $2 \times 2 \times 2$ bins (resulting in a final histogram of 1800 bins). In the classification step $m = 1$, since we only have one training example for each gesture. Finally, for the late fusion, the weight $\alpha = 0.8$ was empirically set. As a pre-processing step, DTW was applied to the sequences in order to segment the gestures.

3.3 Evaluation measurements

For the evaluation of the methods, in the context of gesture recognition, we have used the Levenshtein distance or edit distance, proposed by the ChaLearn challenge. This edit distance between two strings is defined as the minimum number of operations (insertions, substitutions or deletions) needed to transform one string into the other. In our case, the strings contain the gesture labels (from 1 to 5 gestures) detected in a video sequence. For all the experiments, we compute then the mean Levenshtein distance (MLD) over all sequences and batches.

[2]Some examples of gestures from different *lexicons* can be observed in the ChaLearn Gesture challenge URL: `http://gesture.chalearn.org/data/data-examples`

3.4 Results

Table 5 shows a comparison of different state-of-the-art RGB and depth descriptors, including our proposed VFHCRH, using our BoVDW approach. In the case of RGB descriptors, HOF by its own is the one performing the worst. In contrast, the concatenation of HOF to HOG descriptor outperforms the simple HOG. Thus, HOF contributes adding discriminative information that HOG does not have. In a similar way, looking at the depth descriptors, one can see how the concatenation of the CRH to the VFH descriptor clearly improves the performance respect the simpler VFH. The bar plot in Figure 4 shows the performance in all the 20 development batches separately, when using late fusion in order to merge information from the best RGB and depth descriptors (HOGHOF and VFHCRH, respectively). When using late fusion, a MLD of 0.2714 is achieved. Furthermore, we also applied late fusion in a 3-fold way, merging HOG, HOF and VFHCRH descriptors separately. In this case we assigned the weight α to HOG and VFHCRH descriptors (and $1 - \alpha$ to HOF), achieving a MLD of 0.2662.

	Descriptor	**MLD**
RGB	HOG	0.3452
	HOF	0.4144
	HOGHOF	**0.3314**
Depth	VFH	0.4021
	VFHCRH	**0.3064**

Figure 5. Mean Levenshtein distance for RGB and depth descriptors

4 Conclusion

In this paper, we have presented a Bag-of-Visual-and-Depth-Words approach for gesture recognition using multimodal RGB-D images. We have proposed a new depth descriptor VFHCRH, which outperforms the existent VFH, as well as the state-of-the art RGB descriptors. Moreover, we have analysed the effect of the late fusion for the combination of RGB and depth descriptors in the BoVDW. Finally, we have presented a fully-automatic gesture recognition system, using DTW for a prior segmentation of the video sequences, and the BoVDW approach for the classification of each segmented gesture.

References

[1] R. Bogdan, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *ICRA*, Kobe, Japan, 2009.

[2] Chalearn gesture dataset, california, 2011.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV*, pages 1–22, 2004.

[4] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, Sept. 2005.

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, pages 1–8, 2008.

[6] D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *ECML-98, 10th European Conference on Machine Learning*, pages 4–15, 1998.

[7] M. Mirza-Mohammadi, S. Escalera, and P. Radeva. Contextual-guided bag-of-visual-words model for multi-class object categorization. *CAIP*, pages 748–756, 2009.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

[9] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 3212 –3217, may 2009.

[10] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *Proceedings of the 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 2010.

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.