21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

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

Antonio Hernández-Vela[1,2]
ahernandez@cvc.uab.cat

Miguel Ángel Bautista[1,2]
mbautista@ub.edu

Xavier Perez-Sala[2,3]
xavier.perez-sala@upc.edu

Victor Ponce[1,2]
vponce@cvc.uab.cat

Xavier Baró[2,4]
xbaro@uoc.cat

Oriol Pujol[1,2]
oriol@maia.ub.es

Cecilio Angulo[3]
cecilio.angulo@upc.edu

Sergio Escalera[1,2]
sergio@maia.ub.es

[1] *Dept. MAIA, Univ. de Barcelona, Gran Via 585, 08007 Barcelona, Spain.*
[2] *Computer Vision Center, Campus UAB, Edifici O, 08193 Bellaterra, Barcelona, Spain.*
[3] *CETpD, Univ. Politècnica de Catalunya, Rbla. de l'Exposició, 59-69, 08800 Vilanova i la Geltrú, Spain*
[4] *EIMT, Univ. Oberta de Catalunya, Rbla. del Poblenou 156, 08018 Barcelona, Spain*

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. State-of-the-art RGB and depth features, including a new proposed depth descriptor, 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 in comparison to a 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, these words become visual 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, etc.) or detecting points with relevant properties (corners, salient 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].

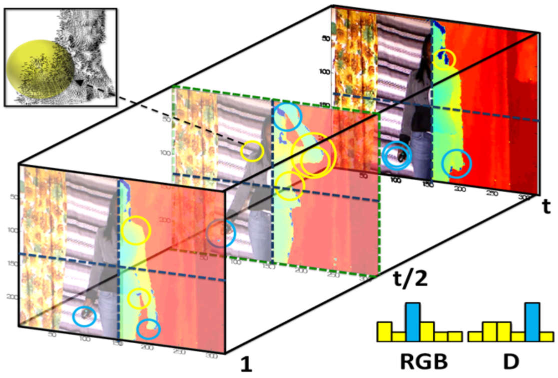

Figure 1. BoVDW approach in a gesture recognition scenario. Interest points in RGB and depth images are depicted as circles. Color of the circles indicates the assignment to a visual word in the shown histogram. Limits of the bins from the spatio-temporal pyramids decomposition are represented by dashed lines in blue and green, respectively. A detailed view of the normals of the depth image is shown in the upper-left corner.

Besides, 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 acquired by a camera, provid-

978-4-9906441-0-9 ©2012 ICPR

ing 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 [9] presented one of the greatest advances in the extraction of the human body pose using RGB-D, 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 [1], which are based on codifying the distribution of normal vectors among regions in the 3D space.

In this paper, we present (a) the 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 (b) combining information of both RGB images and depth maps. We also propose (c) 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 evaluate the presented approach, we (d) 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, (e) 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.

The paper is organized as follows: Sec. 2 presents the BoVDW model for gesture recognition. Sec. 3 presents the results. Finally, Sec. 4 concludes the paper.

2 BoVDW

In this section, we present the BoVDW approach for Gesture Recognition. The BoVDW pipeline is shown in Figure 2 (blue pipeline). Figure 1 contains a conceptual scheme of the approach. The steps of the procedure are described next. Finally, we present the application of the BoVDW to Gesture Recognition (green pipeline in Figure 2).

2.1 Keypoint detection

The first step of BoW-based models consists of selecting a set of points in the image/video with relevant properties. In order to reduce the amount of points in a dense spatio-temporal sampling, 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 searches regions in the image with significant eigenvalues λ_1,λ_2,λ_3 of μ, combining the determinant and the trace of μ:

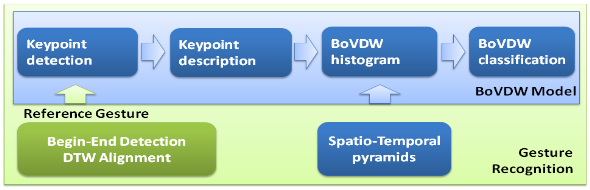

Figure 2. BoVDW-based Gesture Recognition.

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

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 S_{RGB} and S_D.

2.2 Keypoint description

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

2.2.1 VFHCRH

The recently proposed PFH and FPFH descriptors [1] represents each instance in the 3-D cloud of points with a histogram encoding the distribution of the mean curvature around it. Both PFH and FPFH provide $\mathcal{P}6DOF$ (Degrees of Freedom) pose invariant histograms, being $\mathcal{P}$ the number of points in the cloud. Following their principles, VFH describes each cloud of points 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 cloud of points using a different number of descriptors r, where r is the number of stable regions found on the cloud. 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 normal $n^{(i)}$ of the i-th point $p^{(i)}$ onto a plane P_{xy} that is orthogonal to the viewing axis z, the vector between the centroid of the cloud and the camera center, under orthographic projection:

$$n_{xy}^{(i)} = ||n^{(i)}|| \cdot \sin(\phi), \tag{2}$$

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

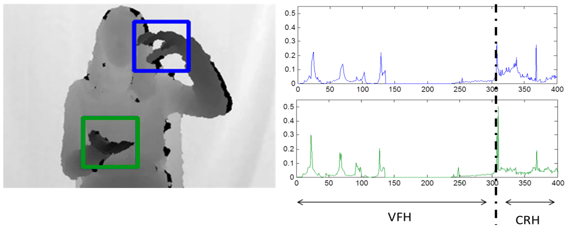

Figure 3. VFHCRH descriptor: Concatenation of VFH and CRH histograms resulting in 400 total bins.

where ϕ is the angle between the normal $n^{(i)}$ and the viewing axis. Finally, the histogram encodes the frequencies of the projected angle ψ between $n_{xy}^{(i)}$ and y-axis, the vertical vector of the camera plane.

In order to avoid descriptors of arbitrary lengths for different point clouds, we describe the whole cloud 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.

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 method –k-means in our case– defines the words from which a query video will be represented, shaped like a histogram h that counts the occurrences 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 are concatenated jointly with the general histogram computed 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 the corresponding histograms, h^{RGB} and h^{D}. Finally, the information given by the different modalities 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 could be used. In our case, we use a simple k-Nearest Neighbor classification, computing the complementary of the histogram intersection as distance:

$$d^F = 1 - \sum_i \min(h^F_{model}(i), h^F_{query}(i)), \qquad (3)$$

where $F \in \{RGB, D\}$. 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:

$$d = (1-\alpha)d^{RGB} + \alpha d^D, \qquad (4)$$

to perform late fusion, where α is a constant relative importance factor.

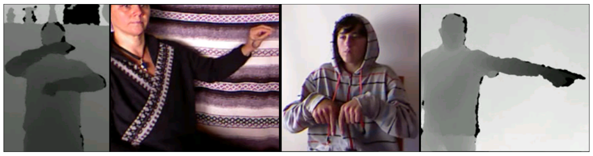

Figure 4. Gesture samples from the ChaLearn data.

2.5 Gesture Recognition system

In order to test the BoVDW model representation, we designed a continuous gesture recognition system. First, DTW 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 above. These steps are also illustrated in the green pipeline shown in Figure 2.

3 Experiments

In order to present the results, we discuss the data, methods, and evaluation metrics for 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 (see Figure 4). The sequences are organized in 20 batches, each one of them including 100 recorded gestures grouped in sequences containing from 1 to 5 gestures, performed by the same user. The gestures from each batch are drawn from a different vocabulary of 8 to 15 unique gestures and 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

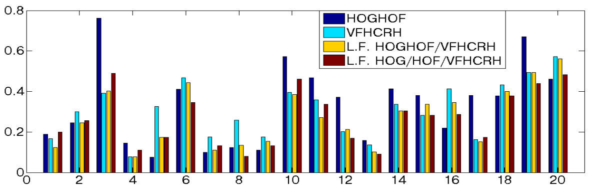

Figure 5. Performance of the best RGB and depth descriptors separately, as well as the 2-fold and 3-fold late fusion of them. X axis represent different batches and Y axis represents the MLD of each batch.

histogram of 1800 bins). In the classification step we use a simple Nearest Neighbor classifier, 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 all 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. 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 detected in a video sequence. For all the comparison, we compute the mean Levenshtein distance (MLD) over all sequences and batches.

3.4 Results

Table 1 shows a comparison between different state-of-the-art RGB and depth descriptors (including our proposed VFHCRH), using our BoVDW approach. In the case of RGB descriptors, HOF alone performs the worst. In contrast, the early concatenation of HOF to HOG descriptor outperforms the simple HOG. Thus, HOF contributes adding discriminative information to HOG. 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 compared to the simpler VFH. The bar plot in Figure 5 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), 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), improving the MLD to 0.2662. From this result we observe that HOGHOF late fusion performs better than HOGHOF early fusion.

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

Table 1. Mean Levenshtein distance for RGB and depth descriptors

4 Conclusion

In this paper, we have presented the BoVDW approach for gesture recognition using multimodal RGB-D images. We have proposed a new depth descriptor VFHCRH, which outperforms VFH. Moreover, we have analysed the effect of the late fusion for the combination of RGB and depth descriptors in the BoVDW, obtaining better performance in comparison to early fusion. 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.

Acknowledgements

This work has been partially supported by the Catalan government and the following projects: IMSERSO-Ministerio de Sanidad 2011 Ref. MEDIMINDER, RECERCAIXA 2011 Ref. REMEDI, TIN2009-14404-C02, TIN2011-28854-C03-01 and CONSOLIDER-INGENIO CSD 2007-00018. The work of Antonio is supported by an FPU fellowship from the Spanish government.

References

[1] R. Bogdan, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *ICRA*, 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): The independence assumption in information retrieval. *ECML*, 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] J. Shotton, A. Fitzgibbon, and M. Cook. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.