Head Modeling Using Multiple-views

Christian Pieringer
Department of Computer Science
Pontificia Universidad Católica de Chile
Santiago
Email: cppierin@uc.cl

Domingo Mery
Department of Computer Science
Pontificia Universidad Católica de Chile
Santiago
Email: dmery@ing.puc.cl

Alvaro Soto
Department of Computer Science
Pontificia Universidad Católica de Chile
Santiago
Email: asoto@ing.puc.cl

***Abstract*—Object detection has attracted great interest of researchers in the computer vision community. Although machine learning approaches has been successful in this task, there are still significant challenges to solve in order to achieve data association, and including information from various points of views. We propose a multiple-view classification approach to bring a gap between advances in machine learning based object detection and multiple view geometry. The key idea is to classify an image sequence of corresponding parts of an object. This scheme allows us to solve problems related to correspondence throughout cameras, and to enhance the detection models with compounded features. This article describes our approach applied in human head modeling by integration of visual information. The experiments demonstrate that our technique improves 2D state-of-art classifiers, using same training conditions. These results are promising and show that our approach can be use effectively to detect objects using multiple views.**

I. Introduction

Object detection and recognition have been relevant research areas in computer vision along the last decade. The most relevant approaches based on machine learning categorize different kinds of objects using visual features extracted from image patches [1]–[3]. These researches focus on monocular scheme, and only few researchers have dedicated to exploit the use of multiple views to improve their performances. A few recent works focus in to demonstrate that 3D information, improve the detection. However, most of them include additional hardware, such as stereo cameras or depth cameras, due to they are focus on mobile robots.

In general, 3D recognition from 2D images is a complex task due to the infinite number of points of views and different illumination conditions [4]. A simple recognition strategy consists in to performed by matching its invariant features with the features of a model. However, it may fail when objects have a large intra-class variation. In [5], a novel representation for 3D objects is presented based on local affine-invariant image descriptors and multi-view spatial constraints. The algorithm exploits the idea that smooth surfaces are always planar in the small. Thus, the matching and then the recognition is possible using photometric and geometric consistency constraints. A disadvantage is its poor performance on texture images. In [6], a similar method based on the relationships among multiple model views enforces global geometric constraints in order to achieve 3D reconstruction from multiple views to recognizing single objects. A disadvantage is its poor performance on non textured images and uniform objects. In [7], a tracking algorithm classifies the head pose base on the low resolution data in a multi-view camera system. An ellipsoid represents the head position and rotation, and a probabilistic framework joint the scores of the individual views. Finally, the tracking algorithm identify the real pose. There also is a different path for 3D object categorization, which use combination of information from multiple poses or points of view [8]–[11]. However, they still are mono-focal classifiers. Although, 3D object classification and detection have had progress, especially linking features among views in a discriminative learning framework to create multiple view models of objects, there are still challenges to solve in order to improve data association.

We observe that mono-focal approaches for categorization suffer from: *i)* high efforts to improve classification models in only camera, and *ii)* discard available data from different visual sources. On the other hand, wide baseline stereo systems present an unsolved issue related to correspondence matching, where the same object has various poses or variations simultaneously. We propose an approach to categorize objects using simultaneous visual data, where the key idea is to use all the available visual information presents in a multi-view camera system, Fig. 1. The proposed approach offers several promising advantages in object categorization, including the following main contribution of this paper: improving classification performance using models on compounded visual features acquired simultaneously from the multi-view camera system. This framework let us to enrich the data used to train the models. Thus, we are able to include all the visual information in the same model.

This article presents our approach for people head modeling based on integration of visual information in a wide baseline stereo system. The results show our approach improve classification performance in average precision-recall, with a best performance when the algorithm use four cameras. These results are promising and demonstrate our approach can be used effectively to classify objects in multiple-views environments. The rest of the paper is organized as follows: Section II describes the proposed method. Section III provides implementation details, dataset details and main experiments of using our methodology in real images. Finally, Section IV discuss concluding remarks and future avenues of research.

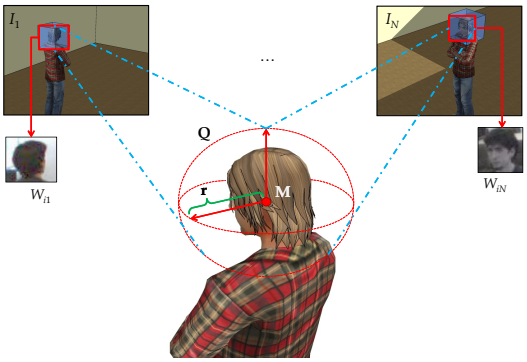


Fig. 1: Diagram of head representation. We use N calibrated cameras $C_1, \ldots, C_N$. In this example we assumes the head is in positions $[X, Y, Z]$. The quadric $\mathbf{Q}$ is projected from 3D space on images $I_1, \ldots, I_N$ to generate the windows $W_1, \ldots, W_N$.

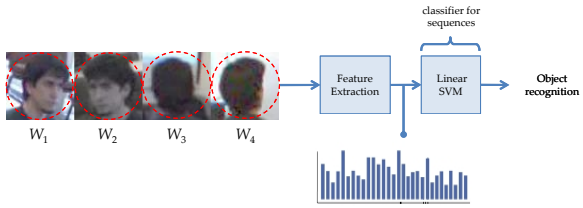


Fig. 2: Block diagram of the proposed method. Our approach includes two main steps: feature extraction, sequence classification. The algorithm begins with a input sequence composed by W_1 to W_N. This sequence represents the quadrics seen from each camera view. We draw the projection Q as dashed red circles to show graphically how to select the maximum parallelepiped subscribed to Q. Each element W_j was cropped, and then rescaled to 64×64 pixels to cope with projection at different size. We extract LBP features for each projection W_j, and finally apply the model in order to classify the sequences.

II. Overview of the Method

Similar to [7], we assumes that an object is represented by an ellipsoid. We use a quadric sphere located at coordinates $\mathbf{M} = [X, Y, Z]$ and radius r, which is totally defined as $\mathbf{Q} = (\mathbf{M}, r)$. We project this quadric onto the images in order to extract bounding-boxes where the object is located. After this process, we get a projected window W_j in the image j, as shown in Fig.2. A set of features represent each projection as inputs for a classifier. We decide over the joint data build up using all the projected windows W_j. More details about quadrics and conics representations can be found in [12].

Our approach requires a fully calibrated multiple view system of N cameras $C_1, \ldots, C_N$, with overlapped fields of view, to compute the geometric model which relates the 3D world homogeneous coordinates $\mathbf{M} = [X\ Y\ Z\ 1]^\mathsf{T}$ to the 2D image coordinates $\mathbf{m}_j = [x_j\ y_j\ 1]^\mathsf{T}$ in each image I_j. This model was obtained for $j = 1, \ldots, N$ cameras using the transformation $\lambda \mathbf{m}_j = \mathbf{P}_j \mathbf{M}$, where λ is an scale factor, and $\mathbf{P}_j$ is the 3×4 calibration matrix of camera C_j [12].

A. Feature Extraction

We rescale each projection W_j to 64×64 pixels to cope with different sizes, and we extract a set of features in pyramidal decomposition for each window [13], [14]. This allows us to represent global and local information from each object instance. Each level $l \in L = \{0, \ldots n\}$ in the pyramid has 4^l cells or patches, and for each cell we compute a descriptor with K bins. The descriptor of the entire image patch W_j has $N_f = K \sum_{l=0}^{L} 4^l$ bins. As recommended in [14], we use $L = 3$.

We use Local Binary Patterns (LBP) proposed in [15] as a measure of texture that uses local appearance descriptors. It is computed comparing a center pixel with its neighbors and this comparison is represented as decimal number. The final LBP descriptor contains $K = 59$ bins. This feature outperform HOG in clutter backgrounds and different textures [16], such as different poses of the head within the patch sequence, as shown Fig. 2.

B. Classifier for Sequences

Once we extracted features on each element W_j, we apply two independent and exclusive scheme each other to classify the projections sequences: gathering features and ensemble of classifiers. Along the experiments, we evaluate both approaches in order to present pros and cons of them. In both cases, we use support vector machines (SVM) with linear kernels as classifiers [17]. SVM with linear kernels improves the classification accuracy and speed andover SVM with no-linear kernels in image categorization problems [18].

As we mentioned previously, both are independent and exclusive each other. A bootstrap strategy, similar as [19], allows to avoid memory overloads and overfit along the training. We let the bootstrap algorithm picks a ratio of misclassified samples and releases a ratio of well classified for the next training round. The next two sections describe how we trained both classification schemes for projections sequences: gathering features and ensemble of classifiers.

1) Gathering Features: In this scheme, we train a single linear SVM classifier with features concatenated as single descriptor, *i.e.* , each W_j have associated a feature vector with N_f bins and the camera system has N_{cam} cameras. Then, the sequence descriptor $N_s = N_{cam} \times N_f$ elements. The key idea is to build up an enriched feature vector, which represent the head structure in a global perspective. After the training, we get a model SVM_{gather} that will be used to classify whole the sequence.

2) Ensemble of Classifiers: The ensemble of classifier is composed by two layers. In the first layer, three individuals linear SVM models, β_1, β_2 and β_3, learn to discriminate among three classes respectively: frontal heads, rear heads, and background. All of them learn in one-vs-all fashion. In the second layer, a new linear SVM model use the scores from the previous layer to discriminate the whole sequence. The three scores, $f_{\beta 1}$, $f_{\beta 2}$, $f_{\beta 3}$ build up the feature vector at the second layer, which has $N_s = N_{scores} \times N_{cam} = 3 \times N_{cam}$

TABLE I: Details of train dataset used for training individual models

class	number of examples
frontal head	916
rear head	916
background	9.583
total	**11.415**

TABLE II: Details of train dataset used for training the ensemble classifer

class	number of examples
frontal head	233
background	3.108
total	**3.441**

elements. This final classifier $\text{SVM}_{ensemble}$ is able to merge the information coming from the camera system.

III. Experiments and Results

In this section, we describe implementation details and results to applying our approach in the classification task.

A. Dataset Details

We build our own multi-view head dataset for training and testing using our camera system, due to the lack of multi-view datasets for people head or people torso. This camera system consists of four synchronized cameras. All images were acquired at 640×480 pixels and 15fps. We manage two train datasets: one for training individual models used for evaluate each projection W_j, and one for training the ensemble.

In both train datasets, we use a set of 10 people placed within a room, spinnig over their Z axis from 0° to 360°. Negative samples include objects such as clothes, computers, walls. We also combined individual samples randomly in order to build artificial negative sequences and enrich the sequence dataset. The test dataset was formed by 300 frames fully labeled from two multi-view video sequences in a classroom or auditorium environment, where people were sat and following a speaker. Sequences are manually labeled in the four cameras. People in this dataset are different to people who appears in the train dataset.

B. Experiments

We evaluate our approach using the both classification schemes. Experiments measure the ability to improve the discriminative power, and centering ability.

1) Enriched features: During the training process, we evaluated the influence of adding information coming from more visual sources. We apply the analysis in terms of classification performance. We started training a classifier only using data from one camera and test this model using the test dataset. We repeated this process along as we add more visual sources. We observe an increase in performance from 40% to 70% of average precision-recall, with maximum considering the four views, as shown Fig 3. As we stated, using more cameras we enhance the features with the complementary information available in the other cameras.

C. Centering

Once we trained both schemes of classifiers, we pick centered and non-centered projections. In Fig. 4, we show three sets of four candidate sequences. The first and second sets, Fig.4a and Fig. 4b belong to head class, and the third set , Fig. 4c belongs to the background class. The classifier for sequences have the main task to discard sequences belong to the background, but we note this model intrinsically also do the task to align the sequence as it learnt in the training process. The higher scores were always given to the best alignments as shown the best scores. All the scores in the third set c) are strongly negative, and therefore all assigned to the background class.

IV. Conclusions

We proposed a head classifier based on wide-baseline stereo camera system. Our approach showed a main contributions of improving classification performance using models on compounded visual features acquired simultaneously from the multi-view system. Both classification schemes show similar behaviors performances. Although we did not address occlusion issues, our experiments showed promising results using information from various points of view in the same scene. The integration of information through our approach is able to codify an structure inherent to the image sequence and therefore an object structure, in this case, head structure. The ensemble also works as a case parts-based approaches, which codify this mid-level structures. One disadvantage is the calibration process, which makes to our approach somewhat rigid to the scene structure. We believe our results are promising, and our approach can be adapted for a another challenging multi-view scenario. For future work, we plan to address occlusion issues and cases where it missing projections within the sequence. We believe possible extract information coded in the sequence which reveal if certain images do not belong to the same window detection. We would like also address the head pose estimation problem using the same framework.

Acknowledgment

This work was supported by the ACT-32 Project focused to improve public transport systems, and Fondecyt N. 1100830 project.

References

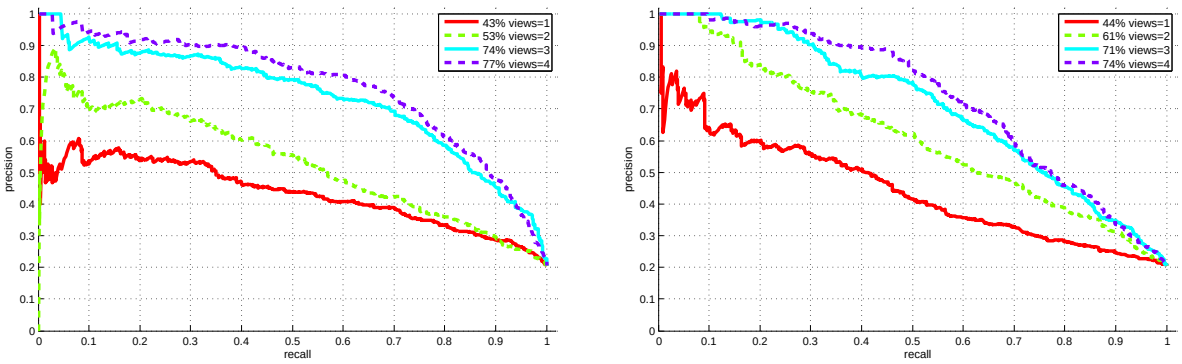
[1] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. CVPR*, vol. 1, pp. 886–893, 2005.

[3] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[4] T. Poggio and S. Edelman, "A network that learns to recognize 3d objects," *Nature*, vol. 343, no. 6255, pp. 263–266, 1990.

[5] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *International Journal of Computer Vision*, vol. 66, no. 3, pp. 231–259, 2006.



(a) (b)

Fig. 3: Both curves show the performance evolution by adding information from various visual sources. Fig. (a) and (b) shows the gather and ensemble strategies, respectively.

(a) (b) (c)

Fig. 4: Image sequences resulting of pass sliding-box among a set of neighbors, and them scores. Box scores are high when the box belongs to head class, and its projections reach the better alignment, as shown in (a)and (b). In (c) we observe background examples and their score, all negatives.

[6] V. Ferrari, T. Tuytelaars, and L. Van Gool, "Simultaneous object recognition and segmentation from single or multiple model views," *International Journal of Computer Vision*, vol. 67, no. 2, pp. 159–188, 2006.

[7] M. Voit and R. Stiefelhagen, "A system for probabilistic joint 3d head tracking and pose estimation in low-resolution, multi-view environments," *Computer Vision Systems*, pp. 415–424, 2009.

[8] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool, "Towards multi-view object class detection," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1589–1596.

[9] S. Savarese and L. Fei-Fei, "3d generic object categorization, localization and pose estimation," *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, 2007.

[10] A. Kushal, C. Schmid, and J. Ponce, "Flexible object models for category-level 3d object recognition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[11] H. Su, M. Sun, L. Fei-Fei, and S. Savarese, "Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories," in *International Conference on Computer Vision*, 2009.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *Proc. CVPR*, vol. 2, no. 2169-2178, p. 1, 2006.

[14] A. Bosch, A. Zisserman, and X. Muñoz, "Image classification using random forests and ferns," *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[15] T. Ojala, M. Pietikainen, and T. Maenpaa, "Gray scale and rotation invariant texture classification with local binary patterns," *Lecture Notes in Computer Science*, vol. 1842, pp. 404–420, 2000.

[16] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Computer Vision, 2009 IEEE 12th International Conference on*, 29 2009-oct. 2 2009, pp. 32 –39.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[18] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 1794 –1801.

[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, 2009.