Automatic Visual Attributes Extraction from Web Offers Images

Angelo Nodari, Ignazio Gallo & Marco Vanetti
Dipartimento di Scienze Teoriche e Applicate, Universita' degli studi dell'Insubria

In this study we propose a method for the automatic extraction of Visual Attributes from images. In particular, our case study concerns the processing of images related to commercial offers in the fashion domain. The proposed method is based on a pre-processing phase in which an object detection algorithm identifies the object of interest, subsequently the visual attributes are extracted using a descriptor based on the Pyramid of Histograms of Orientation Gradients. In order to classify these descriptions, we have trained a discriminative model using a manually annotated dataset of commercial offers, available for comparisons. To increase the performance of the visual attributes extraction, the results provided by the previous step have been refined with an a priori probability which models the occurrence of each visual attribute with a specific product type, opportunely estimated on the dataset.

1 INTRODUCTION

In the recent years we are experiencing a growing interest turned towards visual attributes and their usage in support to many different task: classification, recognition, content-based image retrieval etc...

The concept of visual attribute has been firstly formalized and analyzed by (Ferrari and Zisserman 2007) who propose a generative model for learning simple color and texture attributes from loose annotations and (Farhadi et al. 2009) which learn a richer set of attributes including parts, shape, materials, etc

In the field of face recognition (Kumar et al. 2011) used a set of general and local visual attributes to train a discriminative model which measures the presence, absence, or degree to which an attribute is expressed in images in order to compose a signature of visual attributes. In (Sivic et al. 2006; Anguelov et al. 2007) the main goal consists in finding all the occurrences of a particular person in a sequence of pictures taken over a short period of time. In particular the use of visual attributes to extract information about the hair and clothes of the people has given a consistent contribution in the management of all the cases where the people move around, change their pose and scale, and partially occlude each other.

There are also successful applications of the visual attributes in the field of security and surveillance systems, for example in (Vaquero et al. 2009) the authors used visual attributes instead of the standard face recognition algorithms, which are known to be subject to problems like lighting changes, face pose variation and low-resolution. They search for people by parsing human parts and their attributes, including facial hair, eyewear, clothing color, etc.

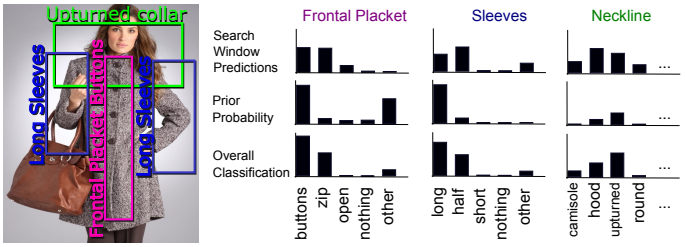

Figure 1: Example of the attribute extraction phase using the proposed method for the three types of attribute analyzed in this study.

In (Wang and Mori 2010) the authors demonstrate that object naming can benefit from inferring attributes of objects and that, in general, the attributes are not independent each other.

The visual attributes express local or general characteristics of a subject, while color, texture and shape are the global features most commonly used, they are also the less interesting in the particularization of the object of interest. They are also used in the aforementioned work and we have analyzed them in a previous work (Gallo et al. 2011), but in this paper we have focused only on the local attributes. These visual attributes are very domain-specific and therefore contain much information that can be used in various fields. In order to be exported to other domains, the extracted attributes must be carefully selected. Moreover, to ensure the applicability of the

Cloth Type	Num	Cloth Type	Num
gown	117	gym suit	57
tunic	108	top	119
tailleur	120	overcoat	118
t-shirt	148	overalls	114
pullover	131	polo	121
padded jacket	130	cloak	111
sweater	120	knitwear	135
raincoat	102	vest	125
short coat	143	jacket	133
sweatshirt	133	turtleneck	102
shrugs	115	cardigan	130
coat	128	camisole	125
shirt	137	blouse	112
blazer	109	short dress	123
dress	202		

Figure 2: Summarization of the types of clothes in the DVA dataset used in this study.

proposed method in real application contexts, the extraction time of the visual features must not increase disproportionately with the number of attributes that have to be extracted and at the same time we want that these algorithms can be very fast.

To the best of our knowledge, the use of the visual attributes in the online shopping has not been exploited yet and for this reason we consider this work very innovative in this area. For example the visual attributes can be used by a user as a method to search the products with particular characteristics which may be congenial for him. For example they can be used in a typical faceted navigation, where the user can search for an item in a structure where all the facets of an item are a possible entry point. At the same time these visual attributes can be integrated in Content-Based Image Retrieval engines for the estimation of the similarity between different images, for example in a query by example system.

2 PROPOSED METHOD

The automatic visual attributes extraction method, proposed in this study, involves the use of a search window whose size and position are in function of the type of visual attribute to search and the bounding box of the object of interest. After have positioned this window, we build a pattern which is then classified by a Support Vector Machine (SVM) (Cortes and Vapnik 1995) in one of the attribute classes which we take into account.

The global features used in the works mentioned in Section 1, such as color and texture, can not be used as visual attributes in this domain because they are not discriminative enough since clothing of the same category may be of different colors and textures. Nevertheless, these attributes have been used successfully in other domains, where the color and texture features are more discriminative such of the case of the dataset of animals (Lampert et al. 2009). For this reason we

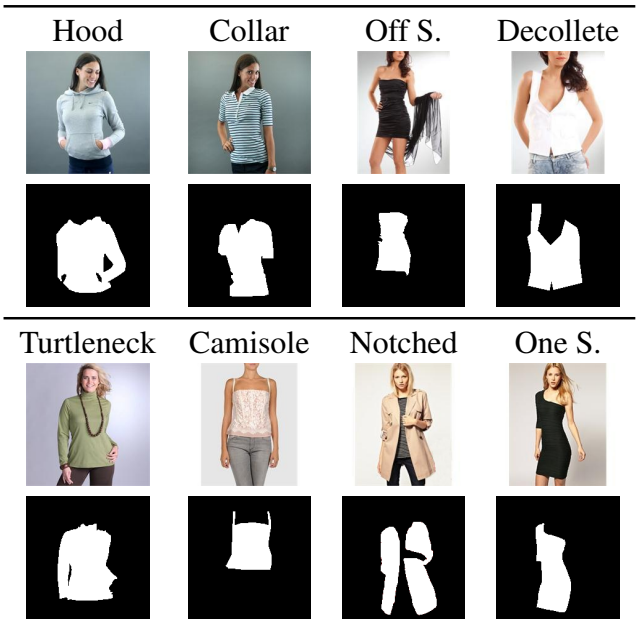

Figure 4: Example of images from the visual attribute Drezzy Dataset grouped by their neckline classes

have chosen more domain specific attributes such as neckline shape, frontal closure and sleeve types.

One problem identified in all the work that extract visual attributes, involves the application of features over the whole image without having first identified, even in a coarse way, the object of interest. Therefore, to properly extract the visual attributes only from the Object of Interest avoiding to be distracted by the background and introducing noise in the extracted data, we necessarily have to distinguish the subject from the background. To perform this step, we relied on our previous work called MNOD (Gallo et al. 2011; Gallo and Nodari 2011) which consists in an algorithm able to perform a segmentation of the object of interest in a specific domain, in this case applied to the fashion domain.

After have detected the object of interest, we select all the manually labeled regions of interest and we estimate the relative position in relation with the bounding box that contains the object of interest. As a result we obtained that to properly look for the neckline attribute we have to position the search window at the upper side of the object of interest, the sleeves attribute in the lateral side and the frontal placket in the middle. This step can be considered trivial, but its formalization allows to easily transfer the entire visual attributes extraction process from the context addressed in this study to any other one.

Once we have identified the location where to place the search window, we resize it in according to the best parameters estimated in Section 3. After that we read the information within the image and represent it as a PHOG (Bosch et al. 2007) features into a pattern that is classified by an SVM in one of the classes of the visual attributes which we are looking for.

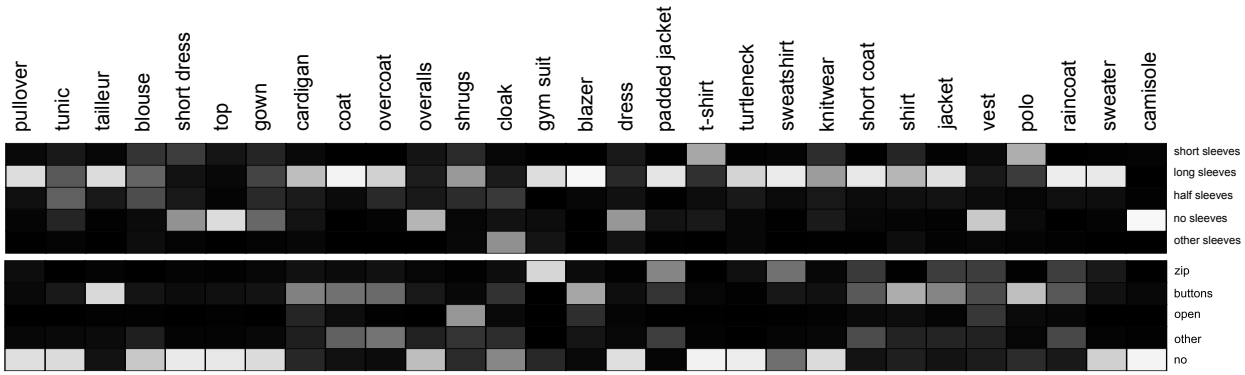

Figure 3: Visual Attribute Signatures for the Sleeves and Frontal Placket Attribute for every product type in the DVA dataset. The gray level represent the probability that a visual attribute may appear associated to a specific product type. To a gray value very high, corresponds a high probability that the pair Attribute Visual and type of product appears and vice versa.

One of the objectives of this study consists in showing how the use of the visual attributes occurrence information, depending on the type of product, can be used in support of the attributes classification phase. This type of use of the attributes has already been addressed in earlier papers such as, for example (Lampert et al. 2009), who have used, however, high-level of attributes manually associated to each image in support of the classification of objects. Instead, in this study, the visual attributes are fully automatically extracted from the image, because in our domain there are not always manual annotations and we want to focus and analyze only the extraction of information from images. The probability distribution of the visual attributes, depending on the types of product, was estimated on the training set and showed according to the representation of the Class Attribute Matrix (Kemp et al. 2006) in Figure 6 and 3.

One of the most difficult task is to place the window for the search of the visual attributes in the correct position. In the experimental phase we have automatically calculated the best position of the window relative to the bounding box of the object of interest. In spite of this, a variation in the positioning of the window is reflected on the accuracy in the extraction phase of the attributes. To overcome this problem, the window is moved arbitrarily respect to its position in order to obtain k readings. For each of these readings is then performed a prediction using the trained SVM. In this way we obtain a set of predictions which are integrated with the values of the a priori probability in order to obtain a result on the extraction of the visual attribute, which minimizes the risk of committing an error. In the experimental section is shown how this method increases the performance in the extraction of the visual attributes.

Given T, the set of product types, and A, the set of visual attributes, the function $s : T \times A- > \mathbb{N}$ counts the number of occurrences of an attribute $a_i \epsilon A$ given the product type $t \epsilon T$. The function $\widehat{f} : T \times A- > \mathbb{R}$ returns the probability estimated on the training set, that a specific attribute $a_i \epsilon A$ may occurs given a product type $t \epsilon T$. In particular $\widehat{f}$ is computed as

$$\widehat{f}(t, a_k) = \frac{s(t, a_k)}{\sum_{i=0}^{n} s(t, a_i)}$$

where n is the number of types of attribute in A. This information is combined with the predictions of the SVM in order to minimize the classification error.

Given p a pattern from the set of patterns P and an observed attribute a, the prediction function $n : P \times A- > \mathbb{N}$ returns the number of predictions of the class attribute a for the pattern p for all the readings of the search window. The predicted attribute class c is defined as follows

$$c = \arg\max_i(\max(\epsilon, f(p, a_i)) \cdot n(p, a_i))$$

where ϵ is a constant to avoid the 0 valued case, situation which corresponds to the probability that a particular attribute a doesn't occur associated with a product name p.

3 EXPERIMENTS

In order to evaluate the proposed method we have built a dataset consisting of 3523 images, and for each of them we have manually labeled the visual attributes analyzed in this study. Because our case of study consists in a set of images related to the fashion domain, they were downloaded from an internet web site for the shopping online[1]. The dataset collected was uploaded to our homepage in order to be used by other methods for comparison[2] and it is named *Drezzy Visual Attribute (DVA)* Dataset.

[1] `http://www.drezzy.com`

[2] `http://www.dicom.uninsubria.it/arteLab/`

The collected dataset is composed by different types of clothing and, to reduce the variability in this domain, we focused on the woman clothing worn in the upper part of the body, whose images are characterized by a consistency in the appearance of clothes and human poses. The dataset consists of over 3,000 images divided in 29 different product types associated to different sets of visual attributes such as 16 classes of neckline shape, 5 classes of sleeves type and 5 classes of frontal closure type. All the product types are summarized in Figure 2 and a set of example images of the *Neckline* attribute are shown in Figure 4. The same approach followed in this domain can be simply adapted and used in other contexts. These images have a resolution of 200x200 pixels because they came from the online domain where there is a constraint on the size of the image and a consistent JPEG compression. These factors significantly complicate the extraction of the visual attributes, because at this resolution is very difficult to extract this kind of information.

3.1 Windows Parameters

To select where to place the search window, we estimated on the training set all the manually placed bounding box. Instead to select the best size of the search window, we performed an experiment varying the proportions of width and height of the window and observing the Kappa value (Cohen 1960), estimated on the DVA dataset. The results have shown that the variation of the observation window, with the exclusion of proportions that lead to degenerate dimensions, it is not found significant changes.

3.2 Features Configuration

We tried different features applied to the problem of the visual attributes extraction. Features such as color and texture were discarded for the reasons discussed in Section 1. In this context we have tried to apply the state of the art in the extraction of local information from the images using the Bag of Visual Words (BOVW) (Yang et al. 2007; Chen et al. 2009; Csurka et al. 2004). We used the SURF algorithm as a feature descriptor and we have performed a trial and error test to select 150 as the best number of words for the BOVW dictionary. The performance in the extraction of visual attributes estimated on the DVA dataset corresponds to a $Kappa = 0,14$. This result is consistent with the analysis performed on the DVA dataset, which shows how an excessive lossy compression of the images leads to a reduction in the quality and therefore the possibility to extract local information, fundamental to discriminate on the different visual attributes. Therefore the choice of another feature that can capture this type of information, working also with degraded images, fell on the PHOG feature.

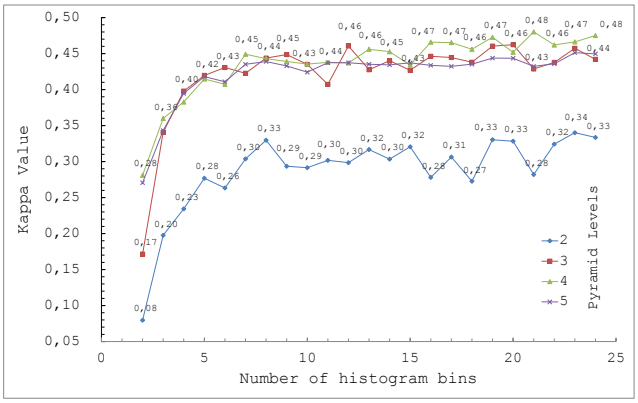

Figure 5: Experimental results of the PHOG parameters tuning using the DVA Dataset. There are plotted 4 series which correspond to the number of pyramid levels, on the horizontal axis are reported the Number of Histogram Bins in function of the vertical axis corresponding to the Kappa value.

In order to select the best PHOG parameters we have performed a tuning experiment on the DVA dataset using a fixed position and a fixed size of the reading window relative to the bounding box of the object of interest. We have evaluated the number of histogram bins and the number of pyramid levels in according to the Kappa value estimated on the dataset, the result are shown in Figure 5. Considering the computational time and results in figure, a good balance in setting the parameters corresponds to set the number of histogram bins to 12 and the number of pyramid levels to 2.

3.3 Visual Attributes Extraction

We have evaluated the extraction of three different visual attributes, with the method proposed in this study, using the DVA dataset and for each visual attribute we discuss about the performance results.

The results on the Sleeves Visual Attribute are shown in Table 2 and it is possible to notice how the *Other Sleeves* class is very difficult to predict. This is due to the few number of examples in the dataset and they can be considered as outlayers in the classification task and so this class can be omitted. The result on the *Half Sleeves* class is very low because, as showed in Table 2, it mingles with the class *Long Sleeves*. For this reason a deep analysis on the used feature have to be performed in order to integrate this class in a real application.

The results on the Frontal Placket Attribute are shown in Table 3 and also for this attribute the *Other* class is troublesome for the same aforementioned reasons.

As regards the visual attribute *Neckline* the results are reported in Table 4 where the difficulty of clas-

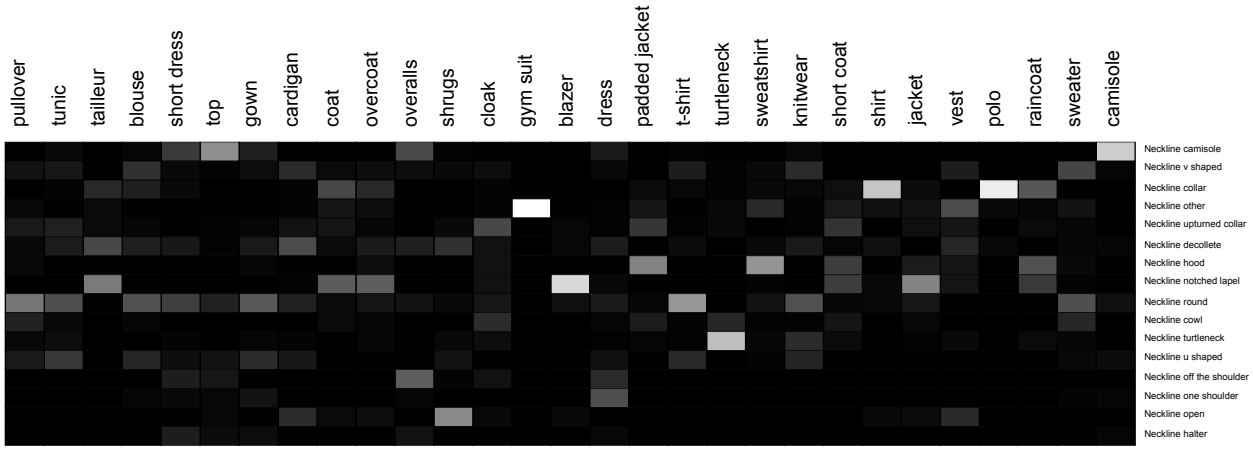

Figure 6: Visual Attribute Signatures for the Neckline Attribute for every product type in the DVA dataset. The gray level represent the probability that a visual attribute may appear associated to a specific product type. To a gray value very high, corresponds a high probability that the pair Attribute Visual and type of product appears and vice versa.

Table 1: Comparison on the application of the estimation distribution of the visual attributes on the DVA dataset for each attribute class

Attribute class	standard method	with estimation
sleeves	0,63	0,66
frontal placket	0,32	0, 43
neckline	0,43	0,59

Table 2: Confusion matrix of the classification of the Sleeves visual attribute with the method proposed in this study.

	short	long	half	nothing	other	CO	PA
short sleeves	28	4	2	4	1	39	71,79%
long sleeves	8	250	20	10	5	293	85,32%
half sleeves	3	5	5	3	4	20	25,00%
nothing sleeves	5	12	4	102	4	127	80,31%
other sleeves	0	0	1	2	1	4	25,00%
TO	44	271	32	121	15	483	
UA	63,64%	92,25%	15,63%	84,30%	6,67%		

Overall Accuracy (OA): 79,92%
Kappa-value: 0,66

sification of this attribute lies in the high number of classes. Nevertheless, it is possible to note how the proposed method is able to correctly recognize a good part of the classes, despite the problem of classification on the generic images of the DVA dataset is very complex.

Each prediction of the discriminative model is combined with the estimated information on the training set, as explained in Section 2, which associates to each product name the probability that it contains a particular attribute. In order to verify if the introduction of this priori information on the estimated training set has given benefits, we performed an evaluation on all the visual attributes with and without this feature showed in Table 1.

The last experiment concerns the calculation of the computational time for the extraction of a visual attribute using a single C# thread, on a Intel®Core ™i5 CPU at 2.30Ghz. The extraction time average of all the elements in the DVA dataset is equal to $304ms$.

Table 3: Confusion matrix of the classification of the Frontal Placket visual attribute with the method proposed in this study.

	zip	buttons	open	other	nothing	CO	PA
zip frontal placket	46	11	4	12	13	86	53,49%
button frontal placket	24	172	9	49	45	299	57,53%
open frontal placket	0	3	23	1	0	27	85,19%
other frontal placket	9	10	5	10	6	40	25,00%
nothing	43	101	23	52	503	722	69,67%
TO	122	297	64	124	567	1174	
UA	37,70%	57,91%	35,94%	8,06%	88,71%		

Overall Accuracy (OA): 64,22%
Kappa-value: 0,43

4 CONCLUSIONS

In this study we have investigated a method for the automatic extraction of visual attributes from images. This technique was applied to the domain of fashion, but as explained in the previous sections, thanks to generic approach that has been adopted, it is possible to extend this method to any other domain assuming to have selected a properly set of attributes on which to work. The experimental results show that estimating the occurrence of the visual attributes on the sample data is of fundamental importance in order to significantly improve the accuracy in the extraction of visual attributes.

Given the lack of available dataset in the domain of the visual attributes extraction, an important contribution led from this work is the introduction of the DVA dataset which can be used for future comparisons.

In a future work, this technique can be extended to other domains, in particular exploring a new method to extract the visual attributes without the information about the type of object present in the image.

REFERENCES

Anguelov, D., K. chih Lee, S. B. Gktrk, B. Sumengen, and R. Inc (2007). Contextual identity recognition in personal photo albums. In *IEEE*

Table 4: Confusion matrix of the classification of the Neckline visual attributes with the method proposed in this study. The labels correspond respectively to the neckline categories: camisole (Ca.), v shaped (V.S.), collar(Co.), upturned collar (Up.), other (Ot.), decollete (De.), hood (Ho.), notched lapel (N.Lap.), round (Ro.), Cowlneckline (Cowl), turtleneck (Turt.), u shaped (U.S.), off the shoulder (off.), one shoulder (One.), open (Open), halter (Hal.)

	Ca.	V.S.	Co.	Up.	Ot.	De.	Ho.	N.Lap.	Ro.	Cowl	Turt.	U.S.	Off.	One.	Open	Hal.	CO	PA
camisole	32	0	4	2	2	7	1	3	3	0	1	3	0	2	0	3	63	50,79%
v shaped	0	13	1	1	1	3	1	0	2	1	0	2	0	0	0	0	25	52,00%
collar	0	3	35	3	3	1	1	2	2	1	1	0	0	0	0	0	52	67,31%
upturned collar	0	0	2	10	0	0	1	1	1	2	3	0	0	0	0	0	20	50,00%
other	0	1	0	2	4	1	0	2	0	1	0	0	0	0	2	0	13	30,77%
decollete	2	3	0	2	3	20	1	6	0	0	0	1	0	0	5	0	43	46,51%
hood	0	0	0	1	7	0	28	1	1	1	1	0	0	0	0	0	40	70,00%
notched lapel	0	3	0	1	1	2	3	32	0	0	0	0	0	0	0	0	42	76,19%
round	2	5	4	0	4	0	5	0	60	1	1	8	3	4	0	0	97	61,86%
Cowl	0	1	0	1	0	0	1	0	0	3	0	0	0	0	0	0	6	50,00%
turtleneck	0	0	0	5	0	0	0	0	0	6	19	0	0	0	0	0	30	63,33%
u shaped	0	0	0	0	0	1	0	0	2	1	0	9	0	0	0	0	13	69,23%
off the shoulder	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	15	100,00%
one shoulder	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	8	100,00%
open	0	0	0	0	1	0	0	0	0	0	0	0	0	0	11	0	12	91,67%
halter	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	4	75,00%
TO	36	29	46	28	26	36	42	47	71	17	26	23	18	14	18	6	483	
UA	88,89%	44,83%	76,09%	35,71%	15,38%	55,56%	66,67%	68,09%	84,51%	17,65%	73,08%	39,13%	83,33%	57,14%	61,11%	50,00%		

Overall Accuracy (OA): 62,53%
Kappa-value: 0,59

Conference on In Computer Vision and Pattern Recognition (CVPR).

Bosch, A., A. Zisserman, and X. Muoz (2007). Representing shape with a spatial pyramid kernel. In *Conference on Image and Video Retrieval*.

Chen, X., X. Hu, and X. Shen (2009). Spatial weighting for bag-of-visual-words and its application in content-based image retrieval. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 867–874.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning*, 273–297.

Csurka, G., C. R. Dance, L. Fan, J. Willamowski, and C. Bray (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22.

Farhadi, A., I. Endres, D. Hoiem, and D. Forsyth (2009). Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*.

Ferrari, V. and A. Zisserman (2007, December). Learning visual attributes. In *Advances in Neural Information Processing Systems*.

Gallo, I. and A. Nodari (2011). Learning object detection using multiple neural netwoks. In *VISAP 2011*. INSTICC Press.

Gallo, I., A. Nodari, and M. Vanetti (2011). Object segmentation using multiple neural networks for commercial offers visual search. In *EANN2011 Engineering Applications of Neural Networks*. ACM Press.

Kemp, C., J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda (2006). Learning systems of concepts with an infinite relational model. In *AAAI*.

Kumar, N., A. C. Berg, P. N. Belhumeur, and S. K. Nayar (2011, Oct). Describable visual attributes for face verification and image search. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Special Issue on Real-World Face Recognition*.

Lampert, C. H., H. Nickisch, and S. Harmeling (2009). Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition*, pp. 951–958.

Sivic, J., C. L. Zitnick, and R. Szeliski (2006). Finding people in repeated shots of the same scene. In *Proceedings of the British Machine Vision Conference*.

Vaquero, D. A., R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk (2009). M.: Attribute-based people search in surveillance environments. In *In: IEEE Workshop on Applications of Computer Vision*.

Wang, Y. and G. Mori (2010). A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision*, pp. 155–168.

Yang, J., Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, MIR '07. ACM.