Towards More Natural Social Interactions of Visually Impaired Persons

Sergio Carrato, Gianfranco Fenu, Eric Medvet, Enzo Mumolo,
Felice Andrea Pellegrino, and Giovanni Ramponi

DIA, University of Trieste, Italy
{carrato, fenu, emedvet, mumolo, fapellegrino, ramponi}@units.it

Abstract. We review recent computer vision techniques with reference to the specific goal of assisting the social interactions of a person affected by very severe visual impairment or by total blindness. We consider a scenario in which a sequence of images is acquired and processed by a wearable device, and we focus on the basic tasks of detecting and recognizing people and their facial expression. We review some methodologies of Visual Domain Adaptation that could be employed to adapt existing classification strategies to the specific scenario. We also consider other sources of information that could be exploited to improve the performance of the system.

1 Introduction

The realization of tools for improving the quality of life of blind people has always been a very active topic of both academic research and industry development. Modern ICT technologies have opened new interesting possibilities; for example, the large available communication bandwidth has made feasible a project such as BeMyEyes [1], where volunteers, using a live video connection, help blind people by answering questions they make, e.g., the expiry date printed on some food, or by giving information about the surroundings, so that they can easily move around. Many other initiatives exist, such as FaceSpeaker [2], a wearable face recognition system which can help the social interaction of a blind person by identifying those who are close to him or her, Horus [3], which relies on video to audio information conversion, or vEyes [4], a non-profit organization which aims at boosting the development of simple application-specific tools for the blind. However, the fact that none of these has received universal acceptance in the blind people community, as it instead happened to the Braille systems almost 200 years ago, seems to suggest that they have not been able to provide solutions with a sufficient degree of efficacy, efficiency and ease of use.

In terms of personal interactions, in particular, the everyday life of persons with a visual impairment brings a number of situations in which "naturalness" is affected. The person with disability and his/her interlocutor are aware that many commonly used non-verbal interaction channels are not available; as a consequence, they are compelled to modify their behaviour to partially compensate for this deficiency. Non-verbal communication includes physical movements

(hand and eyes movements, posture, face expressions), the speaker's appearance (clothes, accessories, make-up) and the distance between communicators. Blind people would feel uncomfortable asking other people to report non-verbal information. According to the focus group of the project "Social Interaction Assistant"[5, 6] the most important non-verbal cues that visually impaired people may need to access are the number and the identity of present people, where a specific person is directing his/her attention, hand and body motions, if someone is behaving inappropriately, and the appearance of a person and how it has changed since the last encounter.

A research project has recently started, funded by the University of Trieste and a private donation, aiming to devise user-friendly vision-based techniques that assist the social interaction of a person affected by a very severe visual impairment or a total blindness. In this paper, we review some recent computer vision techniques, and revisit them according to the requirements of our goal. The originality of this contribution is related to the specific application we refer to. We consider a scenario in which a sequence of images is acquired by a wearable device and is processed by a smartphone in real-time. The number, the identity and the apparent emotional state of the present people have to be discovered and communicated (for instance verbally) to the visually impaired person. It seems obvious to rely on the face appearance to recognize people and infer their emotional state; as a consequence, face detection is a necessary preliminary step. Given the limited computational power at disposal, it seems reasonable to use the same face detector for both detecting and counting people (of course, a drawback of the mentioned approach is that people whose face is not visible in the acquired image will not be detected and counted.) Thus, the sequence of processing steps will be (i) face detection, (ii) recognition of each detected face, and (iii) facial expression recognition on each detected face; indeed, steps ii and iii could be performed in parallel. We refer the reader to [7] for a discussion of some relevant low-level vision issues specific to the described scenario. Here, we deal with computer vision tools at a higher abstraction level, namely those related to the use of classifiers (notice that all the mentioned processing steps require the use of some sort of classifier.) In particular, Section 2 discusses some methodologies of Visual Domain Adaptation that could be used to adapt existing classification strategies to the specific scenario. It is focused on face detection and recognition, but the same techniques can be applied to face expression recognition problems, treated in Section 3. Section 4 finally considers other sources of information that could be exploited to improve the performance of the overall system.

2 Visual domain adaptation

When a classifier is tested against data that possess a distribution different from the training data, a performance degradation usually occurs. For instance, a face detector trained to detect adult faces may fail to detect faces of babies and infants [8]. In pattern recognition, *domain adaptation* refers to a number of techniques aimed at mitigating the performance degradation of a classifier when

it is applied to instances belonging to a domain (called *target domain*) different from that employed during the training phase (called *source domain*). A survey of the recent developments of domain adaptation for visual recognition (i.e., *Visual Domain Adaptation, VDA*) is provided in [9]. Basically, VDA methods try to exploit a few (possibly unlabelled) instances $\mathcal{T}$ of the target domain, along with the whole training set $\mathcal{S}$ of the source domain, to train a classifier capable of working in the target domain. A couple of notable exceptions, that do not rely on the knowledge of $\mathcal{S}$, are [10, 8] as we will explain later on.

2.1 Feature augmentation

Feature augmentation is probably the simplest approach to domain adaptation and is based on the seminal work [11]. If N is the number of features of the original domain, a new vector of features of dimension $3N$ is constructed by duplicating the features and stacking up a null vector 0_N of dimension N. More precisely, given $x_i^s \in \mathcal{S}$ and $x_i^t \in \mathcal{T}$, the corresponding augmented feature vectors, that belong to the new training set, will be $[(x_i^s)^T\ (x_i^s)^T\ 0_N^T]^T$ and $[(x_i^t)^T\ 0_N^T\ (x_i^t)^T]^T$. This simple approach, that exploits the commonalities of source and target domains (first block of the augmented feature space) and the specificity of the source (second block) and the target (third block) is shown to be surprisingly effective in [11], where a kernel version is also proposed. To allow the source and target domain to have a different feature dimensionality (respectively N and M), a projection-based approach has been proposed in [12]. Precisely, the extended feature vectors are $[(W_1 x_i^s)^T\ (x_i^s)^T\ 0_M^T]^T$ and $[(W_2 x_i^t)^T\ 0_N^T\ (x_i^t)^T]^T$ where the projection matrices W_1 and W_2 have the same number l of rows, resulting in an augmented feature space of dimension $l + N + M$. The projection matrices are learned during the training of the adapted classifier. Other feature augmentation approaches are [13, 14], where a manifold of intermediate domains is employed, instead of a single augmented feature space. The feature augmentation approach amounts to building a new training set. As a consequence, the methods could be applied, in principle, to any kind of classifier (a new classifier is trained in a standard way based on the new training set) and in particular on cascade classifiers. Hence these methods are attractive for the considered scenario. Cascade classifiers, indeed, are known to be fast and reliable in object detection, see for instance [15].

2.2 Feature transformation

Instead of augmenting the features, other approaches try to learn a suitable transformation from $\mathcal{T}$ to $\mathcal{S}$. In [16], a transformation is sought such that the instances of the target are mapped close to the instances of the same class belonging to the source and far from those of different class. The transformation is found by solving a constrained optimization problem. If $y = Wx^t$ is the linearly transformed instance, the inner product $(x^s)^T y = (x^s)^T W x^t$ may be view as a similarity measure between x^s and the mapped instance. Based on the known labels of x^s and x^t, a proper constraint is added to the optimization problem to

force similarity or dissimilarity. To prevent overfitting, the objective function is a regularizer $r(W)$, precisely $r(W) = \text{trace}(W) - \log\det(W)$. A kernelized version of the approach is provided in [17] to learn nonlinear transformations. Another feature transformation approach is reported in [18], where a transformation from $\mathcal{S}$ to $\mathcal{T}$ is learnt. More precisely, each instance x_i^s belonging to $\mathcal{S}$ undergoes a rigid transformation $Wx_i^s - e_i$ where W is an orthonormal matrix and e_i is a translation term. In other words, all the source instances are rotated by the same amount and translated by a possibly different amount. The matrix W is found by solving an optimization problem whose goal is to satisfy the constraints $Wx_i^s = [x_1^t\ x_2^t \cdots]Z + e_i$, $\forall i$, by keeping small the rank of Z and the error e_i. The idea is that of expressing the transformed source instances as a combination of few target instances. The transformed source data are then mixed to the target data to train a new classifier. As for feature augmentation approach, the method is not classifier-specific.

2.3 Parameter adaptation

In the parameter adaptation approach, a new decision function for the target domain is formulated, based on a proper perturbation of the original one: $\text{sgn}\left(f_{\mathcal{T}}(x)\right) = \text{sgn}\left(f_{\mathcal{S}}(x) + \delta f(x)\right)$. This approach has been pursued in the Support Vector Machine (SVM) context in [19] and in [20]. The main idea is that of formulating an optimization problem (whose decision variable, in the primal formulation, is w) similar to the standard SVM problem for the target domain, with an additional term in the objective function. The additional term is $||w - w_{\mathcal{S}}||^2$ meaning that a parameter w *close to the parameter* $w_{\mathcal{S}}$ *of the source classifier* has to be found. Notice that $\mathcal{S}$ does not need to be known explicitly, being encoded in $w_{\mathcal{S}}$. Despite the simplicity of the approach, the methods are reported to be effective in visual domain adaptation (for instance, in the task of adapting a classifier trained on bicycles to detect motorbikes). The parameter adaptation approach is suitable for those classifiers whose training amounts to finding a proper vector w of parameters, for instance multilayer perceptrons and SVMs. The new classifier is trained in a non-standard way (because of the additional cost $||w - w_{\mathcal{S}}||^2$) hence a specific code for training has to be written (off-the-shelf SVM implementations are not suitable, for instance), this being a possible drawback.

2.4 Cascade-specific methods

In [10, 8] two VDA approaches are proposed that are particularly relevant for the scenario considered in the present paper. The mentioned approaches are structure-specific, in particular they are suitable for cascade classifiers. Cascade classifiers (see for instance [21]) are composed by a sequence of classifiers that usually have increasing complexity. An instance to be classified may be rejected at any stage of the sequence, being thus classified as negative; it is classified as positive only if it passes through all the stages of the sequence. Perhaps, the most known cascade classifier is the celebrated face detector by Viola and Jones

[22], hence the approaches [10, 8] are relevant for the problem at hand. Furthermore, as opposite to the majority of the other domain adaptation methods, they do not require the knowledge of $\mathcal{S}$. In [10], the following idea is pursued: given a binary classifier whose decision function is $\text{sgn}(f(x))$, the smaller is the prediction value $|f(x)|$, the more uncertain is the assigned class. For the face detection problem, the pre-trained classifier will confidently accept unoccluded, well-illuminated faces, and reject many non-face regions in a given image. Hence it will generate large prediction values for these easy acceptances and rejections. The difficult-to-detect faces will produce smaller prediction values. In [10] it is proposed to update the prediction values of the instances with low prediction from the pre-trained classifier by enforcing the smoothness of $f(x)$ (similar instances are encouraged to produce similar prediction values). In other words, a new function $f'(x)$ is found by enforcing $f(x)$ to be smooth *in the new domain*. The authors of [10] report good results in online adaptation (the classifier is adapted to each image, considered as a new domain). In a sense, detected faces in the image "attract" the difficult images toward themselves. Of course, an underlying assumption is that more faces of similar appearance (for instance, under the same lighting condition) appear in the image. In [8] the cascade classifier is adapted off-line, based on few positive instances from the target domain. In brief, the first stages of the pre-trained cascade (those responsible for the rejection of easy-to-reject instances) are replaced by some stages trained from scratch, based on the few target samples. The remaining stages are selected in order to remove those stages that are responsible of false negatives on the target domain. In [8] the approach is applied to the task of detection of baby faces. The reported results show that the adapted cascade outperforms both the original cascade (trained on human adults faces) and a new cascade, trained from scratch based on the few instances of the target domain.

3 Facial Expression Recognition

Facial expression recognition is performed by extracting from the face image, the features connected to facial expressions and by classifying them. It is a challenging problem because in real environments there are variations in illumination and view angle and because there can be occlusions of the face image, like glasses, or long hair. Moreover, face recognition is generally based on a 2-D face imaging while the face is a 3-D object. However, using 3-D representation of the face images only alleviates the problems.

When benchmarking an algorithm it is recommendable to use a standard testing data set in order to directly compare the results. Therefore, similarly to what has happened for the face detection and recognition problems, many researchers dealing with automatic recognition of facial expressions have developed public datasets. The most popular are The Yale Face Database, which contains 165 grayscale images of 15 individuals with different facial expressions, the Cohn-Kanade Facial Expression Database, which contains 2105 digitized images from male and female subjects, and the Japanese Female Facial Expression

(JAFFE) Database, which contains 213 images of 7 facial expressions by 10 Japanese female models.

3.1 Features for facial expression recognition

The features employed for facial expression recognition can be roughly classified in two main categories, namely geometric features and appearance features. Geometric features can be extracted from the shape of the face or from the location of important facial components such as mouth and eyes. Patil et al. [23] suggest the use of active contour model, called snakes, for tracking lips in face images.

Appearance-based methods use some kind of image processing technique on the facial image in order to extract changes in facial appearance. Appearance features include Gabor [24] and Local Binary Pattern (LBP) [25] features. Gabor features are the output of Gabor filters, which are bandpass filters selective for orientation, and LBP features are local image descriptors which label each pixel on the basis of thresholding of its neighborhoods. It turns out that LBP features are able to statistically describe face characteristics. Among Gabor and LBP methods, LBP is the most commonly used technique for facial expression recognition.

3.2 Classifiers for facial expression recognition

SVM are very popular classifiers for facial recognition. For example, Zhao et al. in [26] propose a method based on LBP features and SVM. They achieved 78.57% on the JAFFE database. Using Gabor features and pseudo 2D Hidden Markov Model, He et al. [27], obtained a 96% accuracy on the JAFFE database. Piparsaniyan et al. describe in [28] a facial expression recognition system based on Gabor feature and simple Bayesian discriminating classifier based on principal component analysis (PCA). They obtain an accuracy of 96.7% on the JAFFE database.

3.3 Real time recognition of facial expression

Since in the application described in the present paper facial expression recognition is executed by devices of low computational powers, tipically smartphones, the development of algorithms which are accurate but at the same time require low computational power is particular important. The real-time system described in [29] is based on a Haar cascade face detector, high-level facial shape features generated from facial landmarks, and a SVM classifier with linear kernel. Accuracy results on two different smartphones (Nexus 4 and Galaxy S3) is 77.5% on the extended Cohn-Kanade dataset. In [30] a real-time system for human-robot interaction based on Gabor filter with a set of morphological and convolutional filters to reduce the noise and the light dependence, and a Dynamic Bayesian Network classifier is described. The authors achieved average emotion detection accuracy of about 94% on a dataset developed by the authors themselves.

Novel techniques based on Deep Learning The deep neural network (DNN) is an emerging technology that has recently demonstrated dramatic success in many applications. These structures need a particular way beyond back propagation to learn the weights of the connections, called Deep Learning. In [31] a facial emotion recognizer based on a convolutional neural network with 65 000 neurons and 5 hidden layers is described. With the extended Cohn-Kanade dataset the authors obtain an accuracy of 99.2% while with LBP features and SVM classifier on the same dataset they obtain 93%. The problem of this kind of algorithm is that the training phase is very computationally intensive, and this is the reason why they use a CUDA machine for training the network. The usage of the network is instead quite fast. The execution of the algorithm on a currently available smartphone takes less than 100 ms.

A fully connected Convolutional Neural Network or cascades of more Convolutional Neural Networks gives rise to Deep Convolutional Neural Networks. The deep CNN has been shown to achieve a strong success for image recognition [32].

4 Context-aware techniques

Social interactions among persons occur, in general, in uncontrolled environments, which may result in low quality information available for the aid machinery, e.g., unoptimal pose, illumination, and so on [7]. On the other hand, other kinds of information could be exploited besides image and video acquired by the device. This supplementary information is often referred to as *context*.

Contextual information may be roughly classified in two categories: (i) low-level data which can be acquired by device sensors other than a camera, and (ii) high-level data which can be acquired by other sources such as Online Social Networks (OSN). Several studies have been carried out which focus on how to better solve specific tasks (mainly face recognition) by using contextual information. Here we review some recent and significant works and highlight how their findings can apply in the scenario considered in this paper.

4.1 Low-level contextual data

The vast majority of today smartphones are equipped with sensors which allow the system to obtain an estimate of its location—i.e., the spatial context: the estimate can be obtained from GPS, or, indirectly, from other sensor readings (opportunistic location). Moreover, all devices are able to determine the current time and hence infer the temporal context. The spatial and temporal contexts have been widely used to improve the accuracy of person identification systems: several noteworthy papers that follow this approach are described below.

The authors of [33] show a method for exploiting spatial and temporal context for improving the accuracy of face recognition applied to images captured by smartphone cameras. They report a remarkable improvement in accuracy (+40%) when using contextual information. Interestingly, this also includes the cell ID—a generally unique number used to identify each Base transceiver station

(BTS), to which the smartphone is connected—which can be used to provide a raw estimate of the location even when the GPS sensor does not operate.

In [34], a similar method is proposed to perform person identification within photo collections. The proposed system gathers three types of contextual information which the authors call temporal proximity, spatial proximity, and co-occurrence. Co-occurrence is the information related to the presence of two or more subjects in the same photo: in the cited work, it is obtained from the Bluetooth signal of nearby devices, when available, with respect to the device which captured the image. The authors conclude that it is possible to improve performance by considering contextual information; besides, they find that temporal proximity is more helpful than spatial proximity and co-occurrence.

A more general framework is proposed in [35], where the availability of location and temporal information is incorporated in a system which is experimentally shown to be effective for person identification in photo collections. The experimental evaluation is carried out using actual GPS readings.

Notwithstanding the cited works seem to suggest that great benefit can be obtained from low-level contextual data (namely, location and temporal information) for real-world person identification, we argue that the actual applicability of this finding to the scenario considered in this paper deserves further investigation. Indeed, the works mentioned above focus on the managing of photo collections, a task that can be performed offline. For instance, consider the case in which a photo P_2 exists in a photo collection in which the persons A and B can be easily identified with a content-only technique, since image conditions are optimal (e.g., pose, illumination); if a photo P_1 exists which has been acquired before P_2 and P_1 and P_2 share the context, then the identification of A and B in P_1 could be aided, even in case of nonoptimal image conditions, by using the context. On the other hand, a system aiming at improving naturalness of social interactions of a visually impaired person should be able to precisely identify persons also when the context (in terms of location and time) is "new", and hence no previous contextual information can be exploited: e.g., the impaired user enters a room where some persons have to be identified.

A radically different kind of low-level data which can be used to perform face recognition is the data coming from 3-D and infrared (IR) sensors. Despite it is known that this data can positively affect face recognition effectiveness, we believe that its impact on the scenario considered in this paper is currently limited, since common devices do not include the related sensors yet. We refer the reader to [36] for a recent survey on techniques which build also on 3-D and IR, also known as multimodal face recognition.

Finally, non visual information can be used also in tasks other than person identification. In [37], a method is presented for real-time, user independent classification of emotions from webcam quality video and audio. The authors shows that the classification accuracy can be improved by considering features derived from audio signal, w.r.t. using only those deriving from video signal. We think that this finding could be of interest also for the scenario considered in this paper, despite the fact that visually impaired persons are likely good in

estimating people emotions from audio. In particular, merging information of audio and video could be useful in the learning phase, when the emotion of a "new" subject can be heard by the user who may then make the system associate the subject's current appearance with that emotion.

4.2 High-level contextual data

The ever increasing ubiquitousness of network-enabled devices able to capture images (e.g., smartphones) and the large adoption of OSN web sites as a tool to store and share those images lead to the existence of large-scale knowledge bases about people appearance and social connections. This high-level information made possible, in the recent past, to reach new milestones of effectiveness and/or practicality in the task of person identification from images. The improvements have been obtained mainly in two ways: (i) by exploiting, in the learning phase, the big (possibly labeled) data available in OSN without significant modifications of existing methods, or (ii) by directly leveraging OSN data to refine the outcome of an identification method. It is worth to note, though, that exploiting web-available information for moving the person identification effectiveness beyond the level that a human would achieve autonomously is perceived as controversial and raises privacy issues. For instance, the authors of [38] highlight the implications of the convergence of face recognition technology and increasing online self-disclosure: they perform two experiments to illustrate the ability to identify strangers online (on a dating site where individuals protect their identities by using pseudonyms) and offline (in a public space), based on photos made publicly available on a social network site.

Researchers of Facebook AI Research show in [39] how they obtained a remarkable improvements in face recognition accuracy on unconstrained environments. To this end, the authors revisit the alignment and representation steps of a conventional face recognition pipeline by including a piecewise affine transformation and a deep neural network, which they train on a dataset including four million facial images.

Many approaches have been proposed for exploiting the social context to refine the results of face recognition methods. For instance, in [40], social cues learnt from large collection of annotated photos by means of association rule mining techniques are used to re-rank face recognition output for the input photo, which results in improved face identification performance with marginal computational overhead. The availability in the web of loosely annotated images of persons can be exploited to build unsupervised face recognition systems. Tools with this aim, possibly tailored to specific kinds of images or persons, have been proposed in [41, 42].

Finally, a fully automatic end-to-end system for face augmentation on mobile devices is proposed in [43]: a smartphone user can point his/her device to a person and the system identifies the person and overlays, in near real-time, a box with his/her information. The tracking algorithm runs on the mobile client, while the recognition runs on a server: people information is obtained offline from OSNs. The cited paper shows the feasibility of a face recognition system

which runs on commodity mobile hardware, a proposal which fits the scenario considered in this paper.

5 Conclusions

We have examined a set of computer vision techniques, revisiting them with the aim of devising the core of a vision-based system able to assist the blind in his/her social interactions. We discussed several visual domain adaptation techniques, which aim at reducing the differences between the training and testing images. Moreover, the emerging deep neural network architectures for image and facial expression recognition have been briefly described. Finally, several approaches for better solving specific tasks using contextual information have been discussed.

Further studies will be devoted to the tools needed to transmit the extracted information to the user. Tactile sensations or sound could be used for this purpose. Moreover, together with the Users' Group of our project, we will realize a set of benchmark sequences to test the various system components. Indeed, even if many datasets exist for the study of face detection and recognition techniques (some have been cited above), none of them is suited to our context and goals.

Acknowledgment

This work has been supported by the University of Trieste - Finanziamento di Ateneo per progetti di ricerca scientifica - FRA 2014, and by a private donation in memory of Angelo Soranzo (1939-2012).

References

1. Be My Eyes: Web site. http://www.bemyeyes.org/ (accessed 5 Aug. 2015)
2. FaceSpeaker: Web site. http://www.facespeaker.org/ (accessed 5 Aug. 2015)
3. Horus Technology: Web site. http://horus.technology/en/ (accessed 5 Aug. 2015)
4. vEyes: Web site. http://www.veyes.it/ (accessed 5 Aug. 2015)
5. Krishna, S., Little, G., Black, J., Panchanathan, S.: A wearable face recognition system for individuals with visual impairments. In: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility, ACM (2005) 106–113
6. McDaniel, T., Krishna, S., Balasubramanian, V., Colbry, D., Panchanathan, S.: Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. In: Haptic Audio visual Environments and Games, 2008. HAVE 2008. IEEE International Workshop on, IEEE (2008) 13–18
7. Bonetto, M., Carrato, S., Fenu, G., Medvet, E., Mumolo, E., Pellegrino, F., Ramponi, G.: Image processing issues in a social assistive system for the blind. In: Image and Signal Processing and Analysis (ISPA), 2015 9th International Symposium on. (Sept 2015)
8. Jain, V., Farfade, S.S.: Adapting classification cascades to new domains. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 105–112

9. Patel, V., Gopalan, R., Li, R., Chellappa, R.: Visual domain adaptation: A survey of recent advances. Signal Processing Magazine, IEEE **32**(3) (May 2015) 53–69
10. Jain, V., Learned-Miller, E.: Online domain adaptation of a pre-trained cascade of classifiers. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 577–584
11. Daumé III, H.: Frustratingly easy domain adaptation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,. (2007) 256–263
12. Li, W., Duan, L., Xu, D., Tsang, I.W.: Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(6) (2014) 1134–1148
13. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 999–1006
14. Gopalan, R., Li, R., Chellappa, R.: Unsupervised adaptation across domain shifts by generating intermediate data representations. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(11) (2014) 2288–2302
15. Zhang, C., Zhang, Z.: A survey of recent advances in face detection. Technical report, Tech. rep., Microsoft Research (2010)
16. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Computer Vision–ECCV 2010. Springer (2010) 213–226
17. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1785–1792
18. Jhuo, I.H., Liu, D., Lee, D., Chang, S.F.: Robust visual domain adaptation with low-rank reconstruction. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2168–2175
19. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: Proceedings of the 15th international conference on Multimedia, ACM (2007) 188–197
20. Aytar, Y., Zisserman, A.: Tabula rasa: Model transfer for object category detection. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 2252–2259
21. Dal Col, L., Pellegrino, F.A.: Fast and Accurate Object Detection by Means of Recursive Monomial Feature Elimination and Cascade of SVM. In Fanti, M., Giua, A., eds.: Proceedings of the IEEE Conference on Automation Science and Engineering, Trieste, Italy, Trieste (2011) 304–309
22. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision **57**(2) (2004) 137–154
23. Patil, R., Vineet, S., Mandal, A.S.: Facial expression recognition in image sequences using active shape model and svm. In: Proceedings of the UKSim 5th European Symposium on Computer Modeling and Simulation. (Dec 2011) 16–18
24. Gu, W., Xiang, C., Venkatesh, Y., Huang, D., Lin, H.: Facial expression recognition using radial encoding of local gabor features and classifier synthesis. Pattern Recognition (2012) 80–91
25. Zhang, S., Zhao, X., Lei, B.: Facial expression recognition based on local binary patterns and local fisher discriminant analysis. WSEAS Trans. Signal Process (2012) 21–31
26. Xiaoming, Z., Shiqing, Z.: Facial expression recognition based on local binary patterns and least squares support vector machines. Lecture Notes in Electrical Engineering **140** (2012) 707–712

27. He, L., Wang, X., Yu, C., Wu, K.: Facial expression recognition using embedded hidden markov model. In: IEEE International Conference on Systems, Man and Cybernetics. (2009) 1568 – 1572
28. Piparsaniyan, Y., Sharma, V.K., Mahapatr, K.K.: Robust facial expression recognition using gabor feature and bayesian discriminating classifier. In: Proc. of Int. Conf. on Comm. and Signal Processing. (2014) 538–541
29. Suk, M., Prabhakaran, B.: Real-time facial expression recognition on smartphones. In: Proc. of IEEE Winter Conference on Applications of Computer Vision. (2015) 1054–1059
30. Cid, F., Prado, J., Bustos, P., , Nunez, P.: A real time and robust facial expression recognition and imitation approach for affective human-robot interaction using gabor filtering. In: Proc. of IROS. (2013) 2188–2193
31. Song, I., Kim, H.J., Jeon, P.B.: Deep learning for real-time robust facial expression recognition on a smartphone. In: Proc. of IEEE Int. Conf. on Cons. Electronics. (2014)
32. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
33. Davis, M., Smith, M., Canny, J., Good, N., King, S., Janakiraman, R.: Towards context-aware face recognition. In: Proceedings of the 13th annual ACM international conference on Multimedia, ACM (2005) 483–486
34. O'Hare, N., Smeaton, A.F.: Context-aware person identification in personal photo collections. Multimedia, IEEE Transactions on **11**(2) (2009) 220–228
35. Kapoor, A., Lin, D., Baker, S., Hua, G., Akbarzadeh, A.: How to make face recognition work: The power of modeling context. AAAI Work (2012)
36. Zhou, H., Mian, A., Wei, L., Creighton, D., Hossny, M., Nahavandi, S.: Recent advances on singlemodal and multimodal face recognition: A survey. Human-Machine Systems, IEEE Transactions on **44**(6) (Dec 2014) 701–716
37. Paleari, M., Huet, B., Chellali, R.: Towards multimodal emotion recognition: a new approach. In: Proceedings of the ACM International Conference on Image and Video Retrieval, ACM (2010) 174–181
38. Acquisti, A., Gross, R., Stutzman, F.: Face recognition and privacy in the age of augmented reality. Journal of Privacy and Confidentiality **6**(2) (2014) 1
39. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1701–1708
40. Bharadwaj, S., Vatsa, M., Singh, R.: Aiding face recognition with social context association rule based re-ranking. In: Biometrics (IJCB), 2014 IEEE International Joint Conference on, IEEE (2014) 1–8
41. Medvet, E., Bartoli, A., Davanzo, G., De Lorenzo, A.: Automatic face annotation in news images by mining the web. In: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, IEEE Computer Society (2011) 47–54
42. Wang, D., Hoi, S.C.H., He, Y.: A unified learning framework for auto face annotation by mining web facial images. In: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM (2012) 1392–1401
43. Dantone, M., Bossard, L., Quack, T., Van Gool, L.: Augmented faces. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE (2011) 24–31