Machine Learning Techniques for Identity Document Verification in Uncontrolled Environments: A Case Study

Alejandra Castelblanco(✉), Jesus Solano, Christian Lopez, Esteban Rivera, Lizzy Tengana, and Martín Ochoa

AppGate Inc., Bogotá, Colombia
{alejandra.castelblanco,jesus.solano,christian.lopez,esteban.rivera, lizzy.tengana,martin.ochoa}@appgate.com

Abstract. Distributed (i.e. mobile) enrollment to services such as banking is gaining popularity. In such processes, users are often asked to provide proof of identity by taking a picture of an ID. For this to work securely, it is critical to automatically check basic document features, perform text recognition, among others. Furthermore, challenging contexts might arise, such as various backgrounds, diverse light quality, angles, perspectives, etc. In this paper we present a machine-learning based pipeline to process pictures of documents in such scenarios, that relies on various analysis modules and visual features for verification of document type and legitimacy. We evaluate our approach using identity documents from the Republic of Colombia. As a result, our machine learning background detection method achieved an accuracy of 98.4%, and our authenticity classifier an accuracy of 97.7% and an F1-score of 0.974.

Keywords: Machine learning · Identity document verification

1 Introduction

Due to the popularity of mobile devices and internet connectivity, interest in distributed enrollment or onboarding processes is raising. Such services typically require pictures of identity documents (ID) as part of the identity verification procedure [1]. In some businesses the proof of identity is crucial and, therefore, it is important to have security mechanisms to prevent identity theft in remote ID verification systems. In fact, several challenges should be addressed in order to accept a document as genuine in scalable onboarding systems [2]. First, the system should localize the document and extract relevant information from pictures taken by the users in uncontrolled environments, such as variable backgrounds, angles, and mobile camera qualities. Second, the system should ensure that the input corresponds to the expected document type. Finally, perceptible document forgery should be detected before accepting the document as genuine.

K. M. Figueroa Mora et al. (Eds.): MCPR 2020, LNCS 12088, pp. 271–281, 2020.
https://doi.org/10.1007/978-3-030-49076-8_26

In the literature, multiple approaches tackle some of these issues individually, for instance, methods for document localization [6,7], text recognition [3,5] and visual similarity comparison [8] have been proposed. However, few papers have addressed complete pipelines for identity document verification using machine learning algorithms [3]. Therefore, more evidence of reliable pipelines and features evaluated in a wide range of datasets is required.

In this paper we propose a practical pipeline for identity document acquisition and verification. Our goal is to design a complete pipeline that takes into account the challenges of real-life acquisition and that could be easily extrapolated to many identity document types (i.e. driving licenses, IDs from various countries). Our contributions are twofold: the first one is related to document localization, where we share gained insights of implementing deep learning techniques for background removal. Also, we propose a set of methods that are necessary for pre-processing images from real-life scenarios. The second contribution is an accurate classifier for document verification based on visual pattern features. For this we rely on novel and already published techniques, which we evaluate on a case study. We also review and evaluate the impact of feature combinations in the performance of classification algorithms.

2 Related Work

Identity document verification aims to determine if an input image belongs to an authentic document class and if the document is legitimate. As described by [8], verification can be performed at the level of content consistency, layout resemblance and visual similarity.

Before performing document verification, the document should be localized and processed from the input image. This step, guarantees a standard input for the authenticity verification system. Most previous studies rely on text recognition or image processing to find the document. For instance, in [6] line segmentation and three-based representations were used to detect quadrilateral shapes. Also, the use of Viola-Jones algorithm complemented with a classifier was proposed by [5]. An accuracy of 68.57% was reported for ID vertices detection in the wild in [2]. Text recognition was used by [3] for document localization. Our work stands out from the literature because it combines a deep learning model to remove complex backgrounds, facilitating the document crop and perspective alignment.

Document verification can be performed by analyzing visual similarity. The pixel-wise information of the document image can be synthesized via features. These descriptors contain unique information from different image components (i.e. Luminance, texture, and structure). Methods such as histogram analysis [20], and color coherence [12] complement the information by comparing intensities and spatial regions. Moreover, analysis of local information has been proposed, with methods such as edge-histograms [11], and structural similarity [19].

Furthermore, studies that perform document type classification and feature extraction were found. Simon et al. [16] classified 74 different ID types through

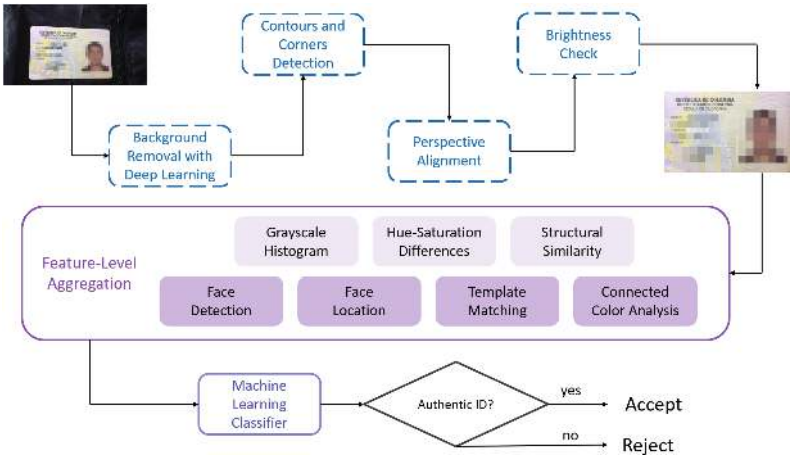


Fig. 1. Pipeline to analyze documents in uncontrolled environments. Blue dashed line boxes depict localization (Module 1). Steps in purple solid line boxes depict authenticity classification (Module 2). (Color figure online)

an SVM. They used a combination of HOG and Color features to gather spatial information achieving a mean class-wise accuracy of 97.7%. Ghanmi et al. [8] performed ID classification with a descriptor based on spatial color distribution, achieving an accuracy of 94.5%. Additional studies that rely heavily on text recognition and deep learning for document classification were found [10,16,18]. Although text extraction is a valuable approach that could complement our proposed pipeline, there is a trade-off, since in-the-wild environment conditions can dramatically impact the performance of the OCR Engines [16] and high resolution images would be required from the end users.

Few papers have addressed complete pipelines for identity document verification in uncontrolled environments. To the best of our knowledge ours is the first work that relies only on visual features to perform all processes (ID localization, classification and verification) in a comprehensive manner.

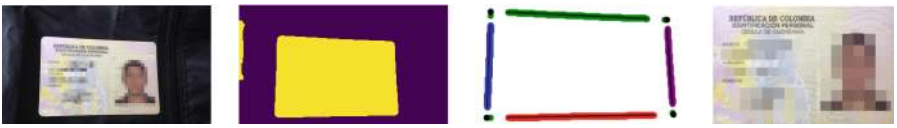
3 Approach

The proposed pipeline for document analysis is divided in two modules, see Fig. 1. The first module addresses the pre-processing requirements of a smartphone document capture in the wild. The second module extracts local and global descriptors of the input image to perform: a) image matching with expected identity document class and b) a basic evaluation of the document authenticity.

3.1 Module 1 - Document Acquisition

Deep Learning Model for Background Removal: We used semantic segmentation to detect the document in the image. This method, consists in the classification

of each pixel into two possible classes, identity document or background. We implemented the UNETS deep learning architecture developed by researchers in [15] to perform pixel classification and build an image with a high contrast background where the document is clearly highlighted.



(a) Original document photo taken by the user.

(b) Identity document mask

(c) Corner detection based on linear regression of contours.

(d) Straightened document after alignment process

Fig. 2. Semantic segmentation followed by perspective alignment and crop.

Crop and Align Perspective: Once the background has been removed, we perform a corner detection analysis, as shown in Fig. 2. First, we find the contour along the document border [17]. Then, the contour is used to perform a linear regression on each side of the document, the four line intersections are defined as the corners of the document. From the selected corners, a geometric transformation matrix is found. The calculated matrix is used to transform the original image into a well oriented document, we used the `warp-perspective` tool from [4]. The sequence of steps is depicted in Fig. 2. We calculated the highest score of template matching, see Sect. 3.2, to detect if the document is upside down.

Brightness Analysis: A brightness analysis is then performed in order to reject images with unsuitable lightning conditions (i.e. flashes). Initially, we separated the image into hue, saturation and value channels (hsv). The third channel, value (v), is used as our measure of brightness. Later, the image is divided into a $n \times m$ bins grid. The brightness value of each bin corresponds to the mean brightness of all pixels which make up that bin. The average brightness (Br_μ) and its standard deviation (Br_σ) are calculated for all bins. A maximum intensity threshold is then computed with Eq. 1, where α controls our *brightness threshold*.

$$Br_{max} = Br_\mu + \alpha \cdot Br_\sigma \tag{1}$$

Following the process, a Gaussian blur filter is applied to reduce noise. Then, each pixel above our given threshold Br_{max} is modified to 255 (white) and below Br_{max} to 0 (black). Afterwards, we group sets of white pixels using a connected component labelling process. These pixel groups are classified as bright zone candidates. Finally, we define a bright spot if the number of pixels in the label is above 2% of the image size.

3.2 Module 2 - Document Verification

The document verification pipeline classifies a set of features that best describe the visual and layout information of the input image. These features should distinguish the original document class from others, and check basic characteristics about the document authenticity. We refer to *global* features as the features that describe the whole image. We call *local* features, the descriptors from a specific region or a particular characteristic in the document which can be adjusted between document types. For this module, we assume that the document is correctly cropped and aligned, and images are resized to (459×297) pixels, this resolution was selected to include some of the lowest picture resolutions available in the smartphone camera market (640×480), keeping document proportions.

Global Features: The first global feature compares the grayscale histograms of the input image against an authentic document image, defined as the ground truth. To handle the variability from the light conditions, the histograms are normalized using a min-max feature scaling. Histogram similarity was measured using the Wasserstein distance (WD). The WD metric proposed by [14], is based on the theory of optimal transport between distributions. WD provided better discrimination between classes, compared to other goodness of fit metrics such as Pearson's chi-squared distance and histogram bin to bin intersection.

The second global feature is generated with a sum of the *hue* and *saturation* differences (HSD) between the input document image X and the ground truth G. For this feature, channels were converted to the HSV space and the document area was split in N rectangular bins, inspired by the bin to bin comparison proposed by [8]. For each bin i, the differences between the average hue $\overline{h}$ and average saturation $\overline{s}$, for X and G were summed. The overall *hue* and *saturation* differences were normalized dividing by the maximum possible differences. The final feature *HSD* was calculated as seen in Eq. 2, with $N = 50$, that is 5 and 10 sections along the height and width respectively.

$$HSD = \frac{\sum_{i=0}^{N} \overline{h}\,(i)_X - \overline{h}\,(i)_G}{179 \cdot N} \cdot \frac{\sum_{i=0}^{N} \overline{s}\,(i)_X - \overline{s}\,(i)_G}{255 \cdot N} \tag{2}$$

The third global feature, structural similarity score (SS), extracts information from the spatial dependencies of the pixel value distributions. For this method, images are compared evaluating functions dependent on the luminance, contrast and value correlations of the pixel arrays, as defined by [19]. This metric compares the structural composition of the background between the input document and the ground truth.

Local Features: Local features are useful to verify the existence of individual elements that are specific to the type of ID document, for instance, pictures, patterns in the background, symbols, bar-codes or headers.

A face within a specific region of the document is represented with two different features. First, a simple 5 point landmark detection was calculated, based on the Histogram of Oriented Gradients and a sliding window. The *Dlib* python

library was used [9]. The output features were: an integer with the number of faces found (NF) on the input image (if all landmarks were found), and a boolean indicating if the face location (FL) matched a predefined valid region.

We used template matching to verify the existence of a specific visual patterns. For this analysis, the input image is analyzed in grayscale, together with an example of the template region from an authentic document. The method consists in sliding a window of the original template over the input image and calculating a correlation measurement. Afterwards, the algorithm returns the coordinates on the input image that has highest correlation with the template. We used *OpenCV* library for the implementation [4].

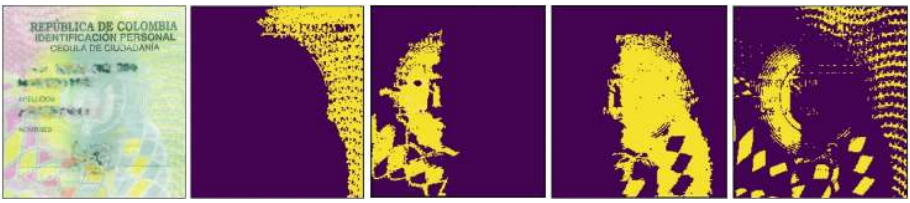


Fig. 3. Color coherence analysis. From left to right: 1) Region analyzed. 2–4) Masks of connected pixels with similar hue values. 5) Mask of connected pixels with similar saturation.

For the Colombian ID case, we chose the document header as template. The correlation measurement that provided better results for discrimination of the authentic document class was the metric *TM-COEFF-NORMED*. The template matching score (TMS) and coordinates of the template location (TML) are exported as features for the classification model.

A variation of the color coherence analysis methods proposed by [8] and [12] was implemented. The proposed method identifies continuous regions with connected color and saturation values and compares these regions between two images. First, the input image is transformed to the HSV space, the hue and saturation channels are discretized in β number of bins. Then, a structural window, that acts as a filter, slides through the discretized image to identify connected color regions, using the tool *label*, from the `ndimage-scipy` python library. Afterwards, connected regions, larger than a certain threshold size, are selected to create binary masks. After applying the described procedure to the hue channel and the saturation channel, a number of N_h hue binary masks and N_s saturation binary masks are created, for both the ground truth image G and an input document image X. To calculate the output features, each mask in G is compared with the closest mask from the image X. For instance, if we are comparing the i^{th} hue mask $M_{huei(G)}$ from image G, the selected mask $M_{huei(X)}$ of image X is the mask with the closest hue value and with the closest euclidean distance to the 2D center of mass from $M_{huei(G)}$. Finally, the Jaccard similarity coefficient between the masks $M_{huei(G)}$ and $M_{huei(X)}$ is the output feature.

In this case study, a section in the front of the Colombian ID cover, that represents a complex pattern with multiple colors was selected, see Fig. 3. For our use case example $\beta = 6$, image G had three binary masks with different hues $N_h = 3$ and one binary mask for saturation $N_s = 1$. The Jaccard similarity coefficients comparing masks in G and X were calculated, thus, the number of color coherence features was four (CC_1, CC_2, CC_3, CS). The binary masks for a Colombian ID document are shown in Fig. 3.

4 Evaluation

4.1 Data Set

The evaluation dataset comprised a total of 101 Colombian identity documents, obtained with the voluntary participants consent. Pictures of the documents were taken by the participants with their own smartphones, without any restrictions on the camera, light conditions or scenario. Image resolutions ranged from (580×370) to (4200×3120) pixels and 40 documents from the collected dataset presented backgrounds with low contrast. For the document verification model (module 2), the machine-learning classifier was tested and trained with features from a subset of 81 identity documents from the collected dataset (positive class). Negative class samples consisted of 40 IDs of other countries and 40 images with multiple environments and patterns. All of them aligned and cropped.

The background removal model (module 1) was built with an augmented dataset. This dataset was generated by applying geometric transformations over different backgrounds. For that purpose, 40 documents were cropped, aligned and placed over 50 different backgrounds. The process was automated through a script, which produced artificial perspectives with specific up-down, right-left tilts in the range of $-15°$ to $15°$ around each axes, and rotations in the range of $[0°, 359°]$. The final result was an augmented dataset consisting of 33382 images. For each document, a ground-truth binary mask was created, representing the desired output. The final images were resized to 128×128 pixels.

4.2 Module 1 - Document Acquisition

Background Removal: We trained a deep neural network using synthetic data augmented from the dataset described in Sect. 4.1.

This dataset was also augmented with empty backgrounds without ID Documents for training, a total of 2254 negative examples, composed of random images and 0-filled masks, were added to the dataset.

For the training, parameters were adjusted to obtain the best performance. The input was tested with both color images and grayscale images; also, a smaller, and more balanced dataset, with only 4766 images was tested. Binary cross entropy (BCE) was used as loss function as it is the default option for binary classification problems, nevertheless, a Jaccard-based loss was also tested.

The best results were obtained after 32 epochs: 98.49% accuracy for the training, 98.41% accuracy for the test, and 0.98 for the Jaccard index for the

test as well. For this model, grayscale images worked better than color images as input. Additionally, the variability of the dataset proved to be more important than the size, since the smaller dataset, generated better results on pictures of ID's over real backgrounds. Finally, BCE showed better results than the Jaccard-based loss. Examples of the outputs from the final neural network configuration in a real world environment can be observed on Fig. 4.

Crop and Align Perspective: The steps of contour and corner detection were evaluated on 96 documents from the real world environment dataset, where the background was removed. In this case, we defined that a crop was successful, with two criteria. First, checking that the score of the template matching analysis was higher than 0.65, which is a good indicator of a correct perspective transformation, and second, by performing visual confirmation of the final result. With those evaluation criteria, we found an accuracy of 88.54%.

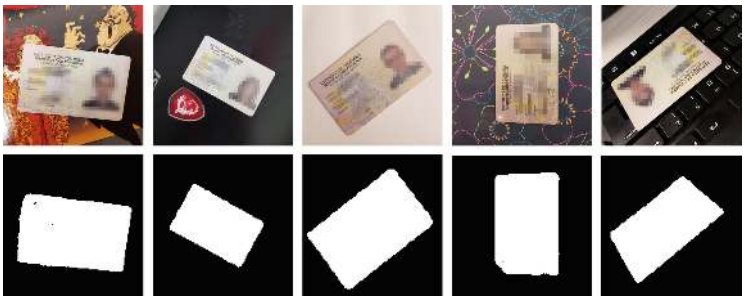


Fig. 4. Document localization process by using image semantic segmentation. Upper images are original pictures and lower images are the deep learning generated masks.

Brightness Analysis: In order to test the brightness analysis, we used a subset of 80 documents, already cropped and aligned, from the original 101 Colombian IDs dataset. We labeled samples with two classes, documents with bright spots (12) and documents without bright spots (68). $\alpha = 2$ was selected to reduce false positives. The proposed method to detect flashes yielded an accuracy of 87.5%.

4.3 Module 2 - Document Verification

Two classification models: Support Vector Machine (SVM) and Random Forest (RF) were tested for document classification, using 11 visual features explained in Sect. 3.2. Features were rescaled to a standard score. The classification tested the features of 81 colombian ID documents, against 80 negative class examples.

Table 1. Document authenticity classification with SVM and random forest.

Features	Accuracy		F1-score	
	SVM	RF	SVM	RF
All (11)	97.5%	97.7%	0.972	0.974
Global (3)	93.0%	90.7%	0.923	0.900
Local (8)	96.7%	97.3%	0.966	0.972

A two-class SVM classifier [13] with RBF kernel was trained. 5-fold cross-validation was used for train-test splits. Using all of the features available, and repeating the train test validation 10 times, an average accuracy of 97.5% with an F1-score of 0.972 was obtained. Classification results with only local or global features can be found on Table 1. In addition, a two-class random forest classifier [13], with gini index as information gain parameter was used. Training with all the features, yielded an average accuracy of 97.77% with an F1-score of 0.974.

Feature contributions for RF are shown on Fig. 5. The results from Table 1 indicate that the visual features selected for document classification are adequate for a production environment. These features can be complemented with content consistency methods or bar-code reading to perform confirmation with official sources.

As depicted Fig. 5, TMS, TML, SS, HSD, and CCS features contributed the most to document verification. We also observed that for the model trained with all the features, three of them explained 90% of the classifier decision.

Additionally, even though the prediction accuracy found when using only global features is 4.5% lower than the SVM model trained with all features, such accuracy is still practical for the proposed verification pipeline. This result encourages to explore the adaptation of global features to other documents, since they do not rely on individual document characteristics.

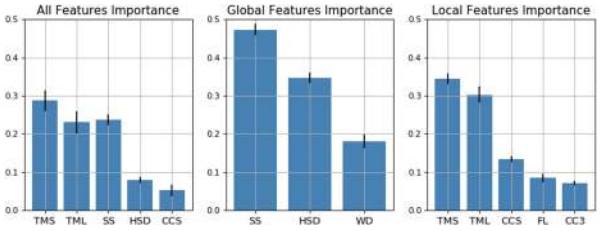


Fig. 5. Feature importance for document classification.

Threats to Validity: Our results indicate that our approach could be a practical and scalable automatic pipeline for remote onboarding processes. The average processing time to execute the document acquisition module was 0.44 s for image sizes of approximately (1200×850) pixels. Document verification takes in average 0.61 s. Additionally, the effort required to adapt the pipeline to other types of documents is expected to be relatively small. However, it would still require a collection of at least 80 authentic documents to train the model, which could be an impediment in many cases due to privacy concerns. For future work it would be interesting to evaluate the performance of the model with fewer training samples. A thorough exploration of the scalability of the approach to different document types is currently missing.

Further explorations on the types of forgery attacks and the degree of the pattern alterations detectable by the classification algorithm are required and could be investigated in future work.

5 Conclusion

A pipeline for identity document analysis was proposed. A module for document acquisition that integrates deep learning for background removal in complex scenarios was formulated and tested. A set of visual features designed for verification of the document type and authenticity were evaluated using machine learning classifiers. Results of this case study show the potential of the methods for complete enrollment processes. In the future we plan to verify if the proposed pipeline can be easily adapted to other document types and larger datasets.

References

1. Arlazarov, V.V., Bulatov, K., Chernov, T., Arlazarov, V.L.: MIDV-500: a dataset for identity documents analysis and recognition on mobile devices in video stream. Comput. Opt. **43**(5), 818–824 (2019)
2. Attivissimo, F., Giaquinto, N., Scarpetta, M., Spadavecchia, M.: An automatic reader of identity documents. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, vol. 2019–10, pp. 3525–3530 (2019)
3. Awal, A.M., Ghanmi, N., Sicre, R., Furon, T.: Complex document classification and localization application on identity document images. In: 2017 14th IAPR ICDAR, pp. 426–431 (2017)
4. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
5. Bulatov, K., Arlazarov, V.V., Chernov, T., Slavin, O., Nikolaev, D.: Smart IDReader: document recognition in video stream. In: 2017 14th IAPR (ICDAR), vol. 6, pp. 39–44. IEEE (2017)
6. Burie, J.C., et al.: ICDAR 2015 competition on smartphone document capture and OCR (SmartDoc). In: 2015 13th (ICDAR), pp. 1161–1165. IEEE (2015)
7. Chazalon, J., et al.: SmartDoc 2017 video capture: mobile document acquisition in video mode. In: 2017 14th IAPR (ICDAR), pp. 11–16. IEEE (2017)
8. Ghanmi, N., Awal, A.M.: A new descriptor for pattern matching: application to identity document verification. In: 2018 13th IAPR International Workshop on Document Analysis Systems, pp. 375–380. IEEE (2018)
9. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
10. Kopeykina, L., Savchenko, A.V.: Automatic privacy detection in scanned document images based on deep neural networks. In: Proceedings RusAutoCon 2019, pp. 1–6 (2019)
11. Park, D., Jeon, Y., Won, C.: Efficient use of local edge histogram descriptor, vol. 2, pp. 51–54 (2000)
12. Pass, G., Zabih, R., Miller, J.: Comparing images using color coherence vectors. In: Proceedings of the Fourth ACM International Conference on Multimedia. ACM (1996)
13. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

14. Ramdas, A., Garcia, N., Cuturi, M.: On Wasserstein two sample testing and related families of nonparametric tests. arXiv:1509.02237 [math, stat] (2015)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. arXiv:1505.04597 [cs] (2015)
16. Simon, M., Rodner, E., Denzler, J.: Fine-grained classification of identity document types with only one example. In: Proceedings of the 14th IAPR, MVA 2015, pp. 126–129 (2015)
17. Suzuki, S., et al.: Topological structural analysis of digitized binary images by border following. Comput. Vis. Graph. Image Process. **30**(1), 32–46 (1985)
18. Wang, J.: Identity authentication on mobile devices using face verification and ID image recognition. Procedia Comput. Sci. **162**, 932–939 (2020)
19. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
20. Van der Weken, D., Nachtegael, M., Kerre, E.: Using similarity measures for histogram comparison. In: Bilgiç, T., De Baets, B., Kaynak, O. (eds.) IFSA 2003. LNCS, vol. 2715, pp. 396–403. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-44967-1_47