Color Seal Extraction from Documents: Robustness through Soft Data Fusion

Aureli Soria-Frisch

Department of Security Technologies, Fraunhofer Institute for Production Systems and Design Technology (Fraunhofer IPK), Pascalstrasse 8-9, 10587 Berlin, Germany
Email: aureli.soria-frisch@ieee.org

Received 11 January 2004; Revised 3 January 2005

This paper presents a framework for the extraction of elements characterized by a particular color hue from color document images. The presented approach attains the detection of official seals, which will be thence analyzed by an embedding system in order to detect possible falsifications. The framework is based on the fusion operator denoted as fuzzy integral, whose robustness with respect to changes in the luminance and the saturation of particular hues is due to the use of the ranking among the input data as an influencing factor in the fusion result. The approach is evaluated on a real data set of tax forms delivered by custom houses, showing its successful performance.

Keywords and phrases: data fusion, fuzzy integral, image segmentation, color processing, document analysis.

1. INTRODUCTION

Offices are one of the human environments rapidly changing due to the evolution of information technologies. Information is abandoning its paper-centered universe to a digital-data-centered one. Administrative, communication, and filing procedures are driven into a digital domain by the ubiquity of different computation facilities. In this context, image processing methodologies for document analysis interface between these two domains, therefore, continuously being challenged by the real world. The here-presented paper takes into consideration an application for the automated analysis of tax forms in custom houses, whose goal is the detection of falsified seals in these documents. In this context, the approach presented in this paper attains the extraction of the seal from the document color image. Once this segmentation methodology has extracted the pixels corresponding to the seal, the image analysis system embedding it proceeds to the detection of its possible falsification. Thus the paper is centered in the analysis of the segmentation stage.

Few methodologies for seal extraction [1, 2, 3] have been hitherto presented in the literature. In [1] the approach is based on the analysis of the seal shape. Since the geometrical aspect of the seals stamped in real offices is extremely variable (i.e., this process cannot be always realized carefully enough), this feature cannot be taken into consideration for the application on hand. Far otherwise the here-presented approach takes the color of the seals as the discriminatory feature in order to segment them from the rest of the document as done in [2, 3]. These approaches attain the full-color segmentation of the document. The segmented images present in this case no special problems. On the other hand, the segmentation approach detailed in this paper attains the segmentation of hand-printed items in a document, which present a high variability in the luminance and the saturation due to a careless stamping process. Furthermore, this approach succeeds in extracting a particular color cluster without fully segmenting the image. This same goal is fulfilled by an application for text extraction on color document images recently presented [4]. Nevertheless, that approach successfully solves the presence of mesh noise in high-quality images, which differs from the seal variability problem formerly described.

The segmentation approach presented here is based on data fusion by considering the color image as a multisensory signal. In multisensory systems, the fusion operator reduces the n-dimensionality introduced in the system by the use of n information sources, for example, the color channels of most usual color models. Thus the fusion operator is a mean for combining the data coming from different sensors into one representational form [5]. A large number of aggregation operators have been developed in the field of soft computing, for example, uninorms [6], OWAs [7], weighted ranking operators [8], and fuzzy integrals [9]. These operators offer a greater flexibility than operators traditionally employed for image fusion.

The fuzzy integral has already been applied to image segmentation [10, 11, 12]. However, the here-presented approach differs from these segmentation procedures, since it
Figure 1: Relational map of different fusion operators. The grade of softness increases in the vertical axis from the top to the bottom and is a result of successive generalizations. In the horizontal axis, the operators are grouped upon their flavor, which defines different families of operators. In the vertical axis, the operators are grouped upon different theoretical frameworks in operations research.

2. SOFT DATA FUSION

Soft data fusion is a conceptual framework presented in [13]. This section gives a general overview on it. The fuzzy integral, which exemplary characterizes this framework, is briefly presented. Furthermore, the reasons of its robustness for image processing are explained.

2.1. Generalization of fusion operators

Fusion operators traditionally used can be considered as hard ones. Furthermore, fuzzy fusion operators were established as generalizations of classical ones (see Figure 1). This mathematical generalization can be considered as a softening process of the operator. The evolution of fuzzy fusion operators from harder to softer ones is based on the inclusion of different parameters influencing the fusion result. Increasing the operator complexity allows as a tradeoff enhancing its robustness, as it is shown in the following paragraphs.

In classical operators, the fusion result exclusively depends on the value being operated on, for example, the result of the sum operator just depends on the summands and thus $1.9 + 3.1$ is always 5. In the weighted operators, for example, weighted sum, an additional factor is taken into consideration, namely, the \textit{a priori} importance of the information sources.

In the theoretical framework of fuzzy logic, the new degree of softness is achieved through the parameterization of the aggregation, for example, T- and S-norms [14], or the consideration of the ranking as a factor upon which the already-mentioned \textit{a priori} importance can be modified. This last strategy is employed in ordered weighted averaging (OWA) operators [7]. Taking into account the ranking of the input data increases the adaptability of the operators and its capability concerning compatibility, partial aggregation, and reinforcement [15].

Fuzzy integrals reflect in the fusion result all the mentioned information: the value delivered by the different sources, their \textit{a priori} importance, and their ranking. The fuzzy integral presents the following positive features in contrast to classical approaches: adaptability, reinforcement capability [15], inclusion of meta-knowledge [15], characterization of the interaction between information sources [16], and tractability of fuzzy information. Moreover, it generalizes both traditionally used fusion operators (e.g., product, sum, minimum, maximum) and other fuzzy fusion operators (e.g., OWAs, weighted ranking operators) [13, 16].

2.2. Soft data fusion through the fuzzy integral

The generalization of classical measures led to the definition of a new type of integrals, which were denoted as fuzzy integrals [9]. This work meant to make more flexible and robust the fusion operation. The fuzzy integral pursues the
approximation of the information binding undertaken by human beings in decision making and subjective evaluation processes. In these processes, different criteria are taken into consideration, weighted and thence joined together in order to generate an answer. Three elements capture the flavor of such a process of bioinspired information fusion, respectively: the linguistic expression of the criteria through fuzzy variables, the weighting through fuzzy measures, and the use of a combination of T- and S-norms [14] as operators.

Fuzzy measure coefficients characterize the a priori importance of the different data in the fusion result. Fuzzy measures generalize classical measures by relaxing the additivity axiom of classical measures, that is, probability measures. Being \( X \) the set of \( n \) information sources, each fuzzy measure coefficient, \( \mu(A_i) \), characterizes the a priori importance of each subset \( A_i \) of \( X \), where \( j = 1, \ldots, 2^n - 1 \). Thus mathematically the fuzzy measures, which are denoted by \( \mu \), are functions on fuzzy sets, \( \mu : \mathcal{P}(X) \rightarrow [0, 1] \), satisfying in the discrete case the following conditions: (1) \( \mu(\emptyset) = 0; \mu(X) = 1 \), and (II) \( A_j \subset A_k \rightarrow \mu(A_j) \leq \mu(A_k) \) for all \( A_j, A_k \in \mathcal{P}(X) \) [16].

There are multiple types of fuzzy integrals, but those known as Sugeno and Choquet fuzzy integrals are the most used ones in applications [16]. The Sugeno fuzzy integral (\( \delta_{\mu} \)) is the generalization of other ranking operators as the weighted minimum or the median and thus presents a combination of the norms minimum (\( \land \)) and maximum (\( \lor \)), whereas the Choquet fuzzy integral (\( C_{\mu} \)) uses a combination of the algebraic product and the addition, becoming a generalization of operators such as the arithmetic mean or the OWAs. The mathematical expressions of these integrals are

\[
\delta_{\mu}(x) = \delta_{\mu}(x_1, \ldots, x_n) = \bigvee_{i=1}^{n} [x_i] \land \mu(A_i)],
\]

\[
C_{\mu}(x) = C_{\mu}(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i \cdot [\mu(A_i)] - \mu(A_{i-1})],
\]

where \( \mu(A(0)) = \mu(\emptyset) \). The enclosed subindices state for the result of a sort operation previous to the aggregation itself, for example, if \( x_1 \geq x_3 \geq x_2 \), then \( x_{(1)} = x_1 \); \( x_{(2)} = x_3 \); \( x_{(3)} = x_2 \). This operation fixes up the coefficients of the fuzzy measures employed in the integration, for example, for the former sorting \( \mu(A_{(1)}) = \mu(x_1), \mu(A_{(2)}) = \mu(x_1, x_3) \), and \( \mu(A_{(3)}) = \mu(x_1, x_2, x_3) \). Therefore, the fuzzy integral defines a different set of weights for each canonical region of the feature hypercube [16], which are defined for the different ranking of the features to be integrated.

From an engineering point of view, it is worth commenting on the robustness of taking the ranking into account. The ranking of the input data is more stable than the value itself, for example, a change in the illumination conditions changes the value of the color values but probably not its ranking relationship. In document analysis, this property applies as well for a change in the stamping pressure, which provokes the aforementioned variability in the luminance and the saturation of ink seals (see Figures 2a and 3a). This robustness is exploited in the here-presented methodology, which is detailed in the following section.

### 3. FRAMEWORK FOR ROBUST COLOR CLUSTER DETECTION

A framework, whose block diagram is depicted in Figure 4, has been implemented for the detection of color clusters. Although the here-presented approach for seal detection is undertaken on an RGB color model, the framework can be applied on any multidimensional one. The used strategy considers the computation of the fuzzy integral on the color channels of the input image with respect to two different fuzzy measures. These are selected in order for the color cluster to be detected to be maximally affected by the change in its coefficients. Thus two gray value images are obtained. Thence the difference image of these results is computed and thresholded in order to generate a binary mask. This mask is dilated and used on the input image in order to segment the seal (see Figure 4). The methodology is formally detailed in the following paragraphs.

Being \( I(x, y) = \{I_d(x, y), I_G(x, y), I_B(x, y)\} \) the input color image in the RGB color model, in the first stage, a difference image \( I_d(x, y) \) is obtained by applying

\[
I_d(x, y) = ||F\mu(x) - F\mu(x, y)||,
\]

where \( F\mu \) states for the images resulting from the computation of the fuzzy integral with respect to the fuzzy measure \( \mu \) on each color pixel, as expressed by (1) or (2) for \( x_1 = I_d(x_1, y), x_2 = I_G(x_1, y), x_3 = I_B(x_1, y) \). A binary image \( I_b(x, y) \) is thence generated by applying a threshold \( \theta \) on \( I_d(x, y) \):

\[
I_b(x, y) = \begin{cases} 
1, & I_d(x, y) \geq \theta, \\
0, & I_d(x, y) < \theta.
\end{cases}
\]

Any binarization procedure based on histogram analysis can be applied for this purpose. In order to get rid of possible failing parts, the mask image \( I_m(x, y) \) results from the application of a once iterated morphological dilation on this image:

\[
I_m(x, y) = I_b(x, y) \circ S,
\]

where \( S \) is a structuring element usually taken as a 4-neighborhood. Finally, the output image \( I_o(x, y) \) is computed by filtering the input image with the obtained mask with a logical AND operator:

\[
I_o(x, y) = I(x, y) \land I_m(x, y).
\]

The obtained results on a first test image show the results of different stages of the framework (see Figure 2). A comparison of the framework performance by applying the Choquet and the Sugeno fuzzy integrals can be undertaken.
Figure 2: Example on the application of the Sugeno fuzzy integral for segmentation of seals on a post letter through the here-presented framework (see Figure 4). (a) Input image. (b) Sugeno fuzzy integral result with the first fuzzy measure, $S_{\mu_1}(x, y)$. (c) Sugeno fuzzy integral result with the second fuzzy measure, $S_{\mu_2}(x, y)$. (d) Final result, $I_0(x, y)$.

Figure 3: Segmentation of seals on a tax form achieved by applying the here-presented framework based on the two types of fuzzy integrals. (a) Input image. (b) Choquet fuzzy integral. (c) Sugeno fuzzy integral.

on Figure 3. The suitability of one or another type of integral is application dependent. Although hitherto no general statements on the selection of the fuzzy integral type can be made [16], our experiments showed a better performance of the Choquet fuzzy integral (compare Figures 3b and 3c).

Lastly, it is worth mentioning how the fuzzy measure coefficients have to be selected. This is undertaken by first determining the canonical region occupied by the color cluster to be extracted. The coefficients that control this canonical region are selected. The value of these coefficients can be set up by a process of extensive search. In this search, a maximal value of $I_d(x, y)$ should be attained when applying these values. The process can be automated by applying numerical optimization procedures, for example, genetic algorithms, although this possibility has not been considered in the here-presented framework.
Figure 4: Block diagram of the here-presented framework for the detection of seals on document images. A fuzzy integral is firstly computed with respect to two different fuzzy measures. The change of fuzzy measure mainly affects the seal color cluster, leaving the other components of the image unmodified. A binary mask image is obtained by subtracting those images, thresholding, and dilating the result. This mask is finally applied in order to extract the seal of the input image.

3.1. Application for color seal detection on document images

In the presented application, which attain the segmentation of seals as the one depicted in Figure 3, the color cluster of the seal is maximally affected by a change in the coefficient \( \mu^i (x_G, x_B) = \mu^i_{GB} \). This fact is a consequence of the position of the bluish color cluster in the canonical region of the color feature space where \( I_R(x, y) \geq I_B(x, y) \geq I_G(x, y) \). Thus the two employed fuzzy measures differ in the coefficient \( \mu_{GB} \). The used strategy exploits in this way the flexibility of the fuzzy integral related to the ranking-based weighting mentioned in the previous section.

In the application on hand, it is suitable to set the coefficients of the first fuzzy measure in order for the fuzzy integral to be equivalent to a minimum operator among the pixels of the color channels. This is achieved by setting \( \mu_{RGB}^1 = 1 \) and the remaining coefficients of \( \mu^i \) to 0. The purpose of this setting is the reduction of parameters. Since \( \mu^1_{GB} = 0 \), the methodology just presents two parameters, that is, \( \mu^1_{GB} \) and \( \theta \).

3.2. Numerical analysis of performance

Due to a nondisclosure agreement with the enterprise delivering the stamped tax-form images, these and the results obtained on them cannot be depicted. Therefore, the segmentation results are commented on hand of an analytical criterion.

Among the criteria presented in [17] the so-called goodness from region shape is selected. Since the seal to be segmented presents a circular shape as the one depicted in Figure 3, the following eccentricity coefficient [18] is computed on the obtained difference images \( I_d(x, y) \):

\[
\varepsilon = \frac{(m_{2,0} - m_{0,2})^2 + 4 \cdot m_{2,1}^3}{(m_{2,0} + m_{0,2})^2},
\]

where \( m_{p,q} \) for all \( p + q = 2 \) are the second-order moments of the gray value image [18]. The real value of \( \varepsilon \), which characterizes shape information, ranges from 0.0 for circular shapes to 1.0 for linear ones. The eccentricity coefficient is rotation, scale, and translation invariant, which compensates for the different position and orientation of the seals in the different images.

4. Analysis of results

The described methodology was applied on 20 documents from the application at hand, which attains the segmentation of the seals for its posterior falsification detection. These documents include a seal that presents all the same motif, which is analogous to the one depicted in Figure 3. The documents are taken from real offices, that is, documents worked out in real offices. Thus they present the seal to be segmented together with different other elements of similar color hue, for example, pen notations, other seals.

The Choquet fuzzy integral outperforms the results of the Sugeno fuzzy integral on the evaluation set due to its smoother response (see Figure 3). The eccentricity coefficient of the difference image computed by applying (3) is computed for different values of \( \mu_{GB} \) and depicted in Figure 5. Because of the circular form of the analyzed seal (see Figure 3), the coefficient should present a value as lowest as possible.

For the sake of comparison, two images with artificial errors are added to the data set. Thus a compact area with the same color hue as the seal was synthetically added on one of the images of the evaluation set. The first disturbing region was placed at 180 pixel distance of the seal center and
The influence of this parameter is not analyzed. Its value can be easily assessed by taking the histogram of \( I_d(x,y) \) into account, since the pixels corresponding to the color cluster build the histogram peak with a highest central value.

## 5. CONCLUSIONS

The paper presents a simple strategy for the extraction of color clusters, where color signals are taken as multisensory ones. Thus the approach is based on the concept of soft data fusion, which is briefly treated in the paper, and uses a fuzzy integral as fusion operator. This fusion operator presents a greater flexibility than operators traditionally used in data fusion applications. The fuzzy integral uses a different set of weights for each canonical region of the hypercube. This dependence of the result of the fusion operator on the ranking of the input data increases the robustness of the operator for image processing applications.

The presented approach is given in the form of a framework, where the fuzzy integral is computed with respect to two different fuzzy measures. The two fuzzy measures are selected in order for the color cluster to be maximally affected by this change. Thus the pixels belonging to the color cluster can be used as a mask in order to extract it from the input image. The procedure, which is characterized by two parameters, is successfully applied for the extraction of color seals from tax forms worked out in real custom houses.

### REFERENCES


**Aureli Soria-Frisch** was born in Barcelona in 1969. He received the “Enginyer Tècnic en Telecomunicaciones” degree (equivalent to the B.S. degree) from the University Ramon Llull (Barcelona) in 1992 and the “Enginyer de Telecomunicació” degree (equivalent to the M.S. degree) from the Politechnical University of Catalonia—UPC (Barcelona) in 1995. Since 1996, he has been at the Department of Security Technologies, the Fraunhofer IPK (Berlin), where he has participated in several research and industrial projects as a Research Scientist and project Leader. He has recently obtained the “Dr.-Ing.” degree (equivalent to the Ph.D. degree) from the Technical University Berlin with a dissertation entitled “Soft data fusion for computer vision,” which describes the application of the fuzzy integral in different industrial systems. He is the author of two journal papers, three book chapters, and several conference papers. His research interest is focused on the following fields: data and multisensory fusion, computational intelligence, soft computing for image processing and analysis, color image processing, texture analysis, and bioinspired image processing.