Looking at Mondrian's Victory Boogie-Woogie: What do I feel?[*]

Andreza Sartori
DISI, University of Trento
Telecom Italia - SKIL Lab
Trento, Italy
andreza.sartori@disi.unitn.it

Yan Yan
DISI, University of Trento, Italy
ADSC, UIUC, Singapore
yan@disi.unitn.it

Gözde Özbal
Fondazione Bruno Kessler
Trento, Italy
gozbalde@gmail.com

Alkim Almila Akdağ Salah
KNAW, e-Humanities Group
Amsterdam, the Netherlands
alelma@ucla.edu

Albert Ali Salah
Boğaziçi University
İstanbul, Turkey
salah@boun.edu.tr

Nicu Sebe
DISI, University of Trento
Trento, Italy
sebe@disi.unitn.it

Abstract

Abstract artists use non-figurative elements (i.e. colours, lines, shapes, and textures) to convey emotions and often rely on the titles of their various compositions to generate (or enhance) an emotional reaction in the audience. Several psychological works observed that the metadata (i.e., titles, description and/or artist statements) associated with paintings increase the understanding and the aesthetic appreciation of artworks. In this paper we explore if the same metadata could facilitate the computational analysis of artworks, and reveal what kind of emotional responses they awake. To this end, we employ computer vision and sentiment analysis to learn statistical patterns associated with positive and negative emotions on abstract paintings. We propose a multimodal approach which combines both visual and metadata features in order to improve the machine performance. In particular, we propose a novel joint flexible Schatten p-norm model which can exploit the sharing patterns between visual and textual information for abstract painting emotion analysis. Moreover, we conduct a qualitative analysis on the cases in which metadata help improving the machine performance.

1 Introduction

Throughout the centuries, Western Art was dominated by representational artworks, which were powerful tools to express feelings and ideas of artists. With the introduction of modern art movements like abstract art, this emphasis shifted and artists explored different means to evoke emotions. For example, abstract artists used the relationship between colour, shapes and textures to convey emotions in a "non-figurative" way: "artists sought to express in their work only internal truths, renouncing in consequence all consideration of external form" [Kandinsky, 1914]. It is therefore intriguing to see if we have the same feelings when we look at these paintings, and if we can train a computer algorithm to report the same.

The relationship between paintings, emotions and art appreciation has been extensively studied in many fields. Psychological studies on aesthetics and emotional responses to art have shown that titles and descriptions influence the perceptual experience of paintings. These works demonstrated that people describe paintings differently after they read the title [Franklin *et al.*, 1993]. They also postulate that title and description of paintings can improve understanding the meaning of the artwork [Leder *et al.*, 2006] and the aesthetic evaluation [Millis, 2001; Hristova *et al.*, 2011]. The influence of title and description is even more crucial for abstract artworks where the visual clues are open to interpretation.

In this paper, we analyze the influence of the metadata (i.e., titles, description and/or artist's statement) associated with the abstract painting and investigate how it can be a significant feature to the automatic emotion recognition of abstract paintings. Specifically, we extract the corresponding metadata for the paintings and apply sentiment analysis to detect the emotional meaning of these texts. We use state-of-the-art computer vision techniques employing colour, shapes and textures to augment computation of the emotional information of the artwork. Finally, we propose two approaches to combine textual features with visual ones: the first approach is based on weighted linear combination, and the second, we propose a novel joint flexible Schatten p-norm model. We apply our multimodal approach on two datasets of abstract paintings: (1) a collection of professional paintings from the MART Museum and (2) a collection of amateur paintings from deviantArt (dA), an online social network site desig-

[*] The title aims to provide an idea about the feeling a title of a painting may induce. The Boogie-Woogie was a cultural movement of music and dance in the late 1920s, and it is characterized by its vivacity, syncopated beat and irreverent approach to melody. This movement fascinated Mondrian as he considered it similar to his own work: "destruction of natural appearance; and construction through continuous opposition of pure means dynamic rhythm." [The Museum of Modern Art, 2015]

nated to user-generated art.

To summarize, our main contributions are: (1) we study the contribution of metadata on positive and negative feelings induced by abstract paintings; (2) we propose a novel joint flexible Schatten p-norm model which can exploit the sharing patterns between visual and textual information for emotion analysis of paintings; (3) we apply our approach on two different datasets of abstract artworks and make a qualitative analysis. This study will contribute to various disciplines and research areas ranging from computer vision (image retrieval) to psychology and aesthetics.

2 Related Work

Several psychological studies provide empirical analysis of how the titles and description of artworks can affect the perception and understanding of paintings. Franklin et al. [1993] showed to the subjects the same painting with two different titles (i.e., the original title and a fabricated title), and asked them to describe the painting. They observed that the change of titles affected the interpretation of artworks. Millis [2001] compared the effects of metaphorical (or elaborative) versus descriptive titles on the understanding and the enjoyment of artworks. He found out that metaphorical titles increase the aesthetic experience more than descriptive titles, or no-titles. In a similar study, Leder et al. [2006] showed that elaborative titles increase the understanding of abstract paintings, but do not affect their appreciation. Hristova et al. [2011] analysed how the style of the paintings and the titles can influence the fixation duration and saccade amplitude of the viewers. When the viewers are presented with the title of a painting while looking at the artwork, the average number of saccades of each viewer to observe the work tend to decline. This finding draws the conclusion that titles lead to a more focal processing of the paintings. The above mentioned research demonstrate that the metadata influence the perception of artworks.

Recently there is an increase on the research focusing on emotions in artworks from the computational perspective. Yanulevskaya et al. [2008] proposed an emotion categorization system for masterpieces based on the assessment of local image statistics, applying supervised learning of emotion categories using Support Vector Machines. Their system was trained on the International Affective Picture System (IAPS), which is a standard emotion evoking image set [Lang *et al.*, 1999], and was applied to a collection of masterpieces. Machajdik & Hanbury [2010] employed low-level features and combined with concepts from psychology and art theory to categorize images and artworks emotionally. They obtained better accuracy in affective categorization of semantically rich images than abstract paintings. Yanulevskaya *et al.* [2012] trained a Bag-of-Visual-Words model to classify abstract paintings in positive or negative emotions. With the backprojection technique the authors determined which parts of the paintings evoke positive or negative emotions. Recently, Zhao et al. [2014] extracted emotion features of images from IAPS dataset of [Lang *et al.*, 1999] based on the principles of art, including balance, emphasis, harmony, variety, gradation, and movement. They concluded that these features can improve the performance of emotion recognition in images. In this work, we use metadata as an additional feature to improve the affective classification of abstract paintings.

Various computational approaches have shown the importance of detecting emotion in textual information [Strapparava and Mihalcea, 2008; Balahur *et al.*, 2011]. Sentiment analysis techniques using computational linguistics to extract the human sentiment in text, is increasingly used in many areas, such as marketing and social networks to improve their products and/or services. Recent works, including [Hwang and Grauman, 2012; Gong *et al.*, 2014], empathize the importance of using textual information associated with images for creating stronger models for image classification. Liu et al. [2011] collected textual information from Internet in the form of tags describing images. They applied two different methods to extract emotional meaning from these tags and combined them with visual features. This study demonstrated how textual information associated with images can improve affective image classification. In our work, we use a multimodal approach to investigate how the title, descriptions and/or artist statements associated with paintings can lead to a stronger model to perform affective analysis of abstract paintings.

3 Datasets and Ground Truth Collection

We conduct our analysis on two sets of abstract paintings, a professional and an amateur one.[1] Both are composed of 500 abstract paintings. In the following subsections we detail the selection process as well as the nature of each dataset.

3.1 MART Dataset: A Dataset of Professional Abstract Paintings

The MART dataset is a set collected in our previous work [Yanulevskaya *et al.*, 2012; Sartori *et al.*, 2015] from the electronic archive of the Museum of Modern and Contemporary Art of Trento and Rovereto (MART). The selected paintings are masterpieces of abstract art, which were created by a total of 78 artists between 1913 and 2008. Most of these artists are Italians, however there are also European and American artists. Some of the artists, such as Wassily Kandinsky, Josef Albers, Paul Klee and Luigi Veronesi, are known for their studies on abstract art in terms of colour, shapes and texture. As most of the artists are from Italy, the titles are in Italian. The available descriptions or the statements of the artists about these paintings are also in Italian.

3.2 deviantArt Dataset: A Dataset of Amateur Abstract Paintings

The amateur collection of abstract paintings has been collected from the deviantArt (dA) website[2], an online social network dedicated to user-generated art. This set was collected in our previous work [Sartori *et al.*, 2015]. dA is one of the largest online art communities with more than 280 million artworks and 30 million registered users. We selected the artworks that were under the category Traditional

[1] The datasets with their respective ground truths are publicly available at: http://disi.unitn.it/∼sartori/datasets/

[2] http://www.deviantart.com/

Art/Paintings/Abstract, and downloaded 8,000 artworks. We used the information reflecting how many times an artwork was favorited as a parameter to downsize the collection from 8000 to 500 paintings[3]. We selected the paintings with the highest and least number of favourites, as well as some paintings randomly in the middle. The collection thus has 500 paintings with 406 different authors. The titles and the artist descriptions in this collection are in English.

3.3 User Study to Analyse Emotions Evoked by Artworks

To collect our ground truth we use the relative score method from our previous work [Sartori *et al.*, 2015] in which we asked people to choose the more positive painting in a pair. The annotation was done online and we provided the following instructions to each annotator: "Which painting in the pair looks more positive to you? Let your instinct guide you and follow your first impression of the paintings."

To annotate and calculate the emotional scores from the paintings we follow the method of TrueSkill ranking system [Herbrich and Graepel, 2006; Moser, 2010]. Developed by Microsoft Research for Xbox Live, the TrueSkill ranking system recognizes and ranks the skills of the players in a game and matches players with similar skills for a new game. With this method the annotation task is more manageable, as it yields a representative annotation with only 3,750 pairs of paintings, instead of 124,750 comparisons $(500 * (500 - 1) * 0.5)$, if each painting is compared with all the remaining paintings in the dataset.

During the annotation process, all paintings are initially considered with the same 'skill' and the painting which is chosen as more positive in a single trial wins a 'game'. Then, the rankings of the compared paintings are updated. Afterwards, the paintings with similar rankings are compared, until each painting is compared with at least 15 other paintings. We consider the results as emotional scores of the paintings, in which lower values correspond to negative feelings and the higher values to positive feelings.

25 subjects (11 females, 14 males) participated in the annotation of the MART dataset. Each person annotated from 19 to 356 pairs of paintings, with a mean of 145. For the annotation of the deviantArt dataset 60 people participated (27 females and 33 males). Each participant annotated from 2 to 436 pairs of paintings, with a mean of 63. The participants were anonymous and they participated in this annotation voluntarily, without getting any reward. There was no time-limit to the annotation procedure: each participant was free to annotate whenever he/she had time to do it.

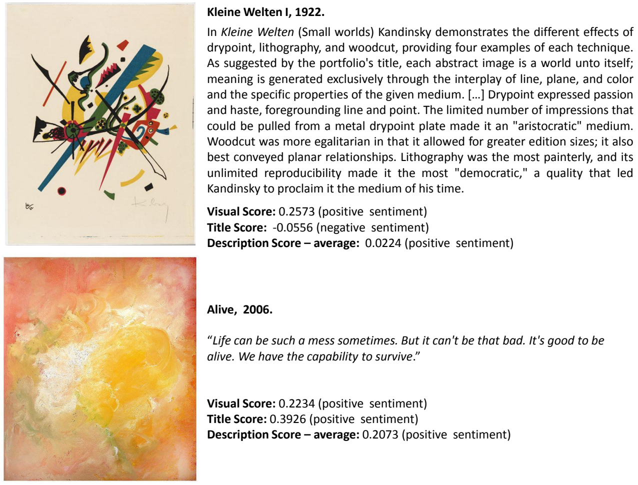
To compose the ground truth of both datasets, a threshold was defined to separate the paintings in positive and negative emotions. As each painting was compared 15 times, we assume that the threshold should be related to one of the paintings chosen 8 times as more positive. Therefore, we consider the TrueSkill ranking values of these paintings, which are the ratings resulted from the comparisons of paintings by using the TrueSkill algorithm. The paintings with TrueSkill ranking equal to or lower than the threshold are defined as negative and paintings with ranking higher than the threshold are defined as positive. For MART dataset, the painting with TrueSkill ranking value equal to 25.1 is used as threshold. In total, we obtain 131 paintings in the negative class and 369 in the positive class. To compose the ground truth for deviantArt dataset, we considered as threshold the painting with TrueSkill ranking equal to 21.0. As a result, 140 paintings were assigned to the negative class and 360 to the positive class.

4 Textual Features for Assessing Emotions in Abstract Paintings

In this section we describe how the sentiment features were extracted from the metadata (i.e., titles, descriptions, artist statement) associated with a painting.

4.1 Text Data Selection

To obtain the descriptions and/or artist statements for MART dataset, we searched through museums and art related websites, as well as the official pages of the artists. In total, we found descriptions and/or artist statements for 158 paintings. Most of these descriptions were in Italian, as the artists were Italian. The titles of the paintings were already provided by the dataset. In deviantArt dataset, the artists themselves provided an opinion and/or a description on/about their paintings. From those, we selected 158 descriptions which represent the artists' intention in detail. The titles and descriptions of this dataset are in English. Figure 1 provides two examples of titles and descriptions for paintings in the MART (top) and deviantArt (bottom) datasets.



Kleine Welten I, 1922.

In *Kleine Welten* (Small worlds) Kandinsky demonstrates the different effects of drypoint, lithography, and woodcut, providing four examples of each technique. As suggested by the portfolio's title, each abstract image is a world unto itself; meaning is generated exclusively through the interplay of line, plane, and color and the specific properties of the given medium. [...] Drypoint expressed passion and haste, foregrounding line and point. The limited number of impressions that could be pulled from a metal drypoint plate made it an "aristocratic" medium. Woodcut was more egalitarian in that it allowed for greater edition sizes; it also best conveyed planar relationships. Lithography was the most painterly, and its unlimited reproducibility made it the most "democratic," a quality that led Kandinsky to proclaim it the medium of his time.

Visual Score: 0.2573 (positive sentiment)
Title Score: -0.0556 (negative sentiment)
Description Score – average: 0.0224 (positive sentiment)

Alive, 2006.

"Life can be such a mess sometimes. But it can't be that bad. It's good to be alive. We have the capability to survive."

Visual Score: 0.2234 (positive sentiment)
Title Score: 0.3926 (positive sentiment)
Description Score – average: 0.2073 (positive sentiment)

Figure 1: Example of titles and descriptions on abstract paintings with their respective emotional scores. The painting on the top is from MART dataset: Vasily Kandinsky, 1922. (*Courtesy of MART photographic archive, Rovereto*) and the description was extracted from The Museum of Modern Art ("MoMA"). The painting on the bottom is from deviantArt dataset: tropicity.deviantart.com/, 2006, (*Courtesy of deviantArt.*)

[3] We selected 500 paintings to construct the deviantArt dataset in order to make an impartial comparison with the MART dataset.

4.2 Applying Sentiment Analysis

To determine the sentiment conveyed by the title and the description of a painting in the deviantArt dataset, we tokenize and part-of-speech (POS) tag the related text with the Stanford Parser [Klein and Manning, 2003], and we use WordNet [Miller, 1995] for lemmatization. For each title or description we simply average the scores coming from SentiWords [Guerini *et al.*, 2013] based on the lemma and POS information. SentiWords is a high coverage resource associating ∼155,000 English lemma-POS pairs to sentiment scores between -1 (very negative) and 1 (very positive).

As (to the best of our knowledge) there is no resource for sentiment scores in Italian, we utilize a multilingual lexical database called MultiWordNet[4] [Artale *et al.*, 1997] for the painting descriptions in the MART dataset. More specifically, we first lemmatize and POS tag the Italian titles and descriptions with TextPro [Pianta *et al.*, 2008]. Then, by using MultiWordNet we find all the English synsets associated with the Italian lemma and POS pairs. We calculate the sentiment score of a lemma-POS pair as the average of the scores provided in SentiWordNet [Baccianella and Sebastiani, 2010] for each synset that it is connected to. SentiWordNet is a lexical resource built for supporting sentiment classification and opinion mining applications and it provides three sentiment scores (i.e., positivity, negativity, objectivity) for each WordNet synset. We consider the difference between the positivity and negativity score as the final sentiment score of a synset. As we did for deviantArt dataset, we average the scores of each lemma-POS pair to determine the overall emotion conveyed by a painting title or description.

5 Fusion of Visual and Textual Features

In this section we describe the classification approach to automatically estimate the emotional valence of a given painting, and provide details on how we used both late fusion and joint flexible Schatten p-norm learning to combine visual and textual features.

5.1 Visual Features

To extract the visual features, we follow our previous work [Yanulevskaya *et al.*, 2012; Sartori *et al.*, 2015]. We use a standard bag-of-words paradigm to extract Colour based LAB Visual Words and Texture Based SIFT Visual Words. We train a Support Vector Machine with a histogram intersection kernel for supervised learning. To test the approach, a 5-fold cross-validation setup is used, where the images are assigned to folds randomly. We run two separate frameworks (Lab descriptors and SIFT descriptors) and then average the scores to combine them.

5.2 Late Fusion with Weighted Linear Combination

To combine visual and textual features we use late fusion with Weighted Linear Combination. To calculate the Weighted Linear Combination we use the following equation:

$$WLC = a * T + (1 - a) * V \tag{1}$$

where T is the sentiment score from text and V is the visual score (the fusion of LAB and SIFT Visual Words) and $a \in [0, 1]$ is a parameter. To choose the parameter a we analyze the performance of the decision fusion (Visual + Text) and set a to the value resulting in the highest mean performance.

5.3 Estimating Paintings' Emotion through Joint Flexible Schatten p-norm Learning

The drawback of late fusion with weighted linear combination approach is that it considers visual and texture information separately before classification. This may not capture well the emotion of the viewers about a painting because the viewers usually judge a painting by looking at it and considering its title and available description at the same time. To better exploit the visual and textual information for the emotion analysis of a painting, we propose a novel joint flexible Schatten p-norm model.

Assuming that $\mathbf{x_i}$ is a d-dimensional feature vector, $\mathbf{X_a} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_{n_a}}\} \in \mathbb{R}^{d \times n_a}$ is the feature matrix from visual information and $\mathbf{X_b} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_{n_b}}\} \in \mathbb{R}^{d \times n_b}$ is the corresponding feature matrix from textual information. $\mathbf{Y_a} = \{\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_{n_a}}\} \in \mathbb{R}^{m \times n_a}$ and $\mathbf{Y_b} = \{\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_{n_b}}\} \in \mathbb{R}^{m \times n_b}$ are label information matrices with m classes. $\mathbf{y_i} = [0, ..., 0, 1, 0, ..., 0]$ (the position of non-zero element indicates the emotion class). To extract the painting's emotion both from visual and texture information, we propose the following optimization problem:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{V_a}, \mathbf{V_b}} & \|\mathbf{Y_a} - (\mathbf{U} + \mathbf{V_a})\mathbf{X_a}\|_F^2 + \|\mathbf{Y_b} - (\mathbf{U} + \mathbf{V_b})\mathbf{X_b}\|_F^2 \\ & + \lambda_1 \|\mathbf{U}\|_F^2 + \lambda_2 \, rank(\mathbf{V_a}, \, \mathbf{V_b}) \end{aligned} \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{m \times d}$ are the common patterns shared across visual and textual information. $\mathbf{V_a} \in \mathbb{R}^{m \times d}$ and $\mathbf{V_b} \in \mathbb{R}^{m \times d}$ are the individual patterns for visual and textual information respectively. In Eqn.(2), the first two terms are loss functions and the regularization term $\|\mathbf{U}\|_F^2$ is used for preventing overfitting. One way to capture the relationship is to constrain the models from both visual and textual information to share a low-dimensional subspace resulting in the rank minimization problem. The last term of rank regularization captures the correlation of the visual and textual information. However, the rank minimization problem is in general NP-hard. One popular approach is to replace the rank function by the trace norm (or nuclear norm).

Inspired by [Nie *et al.*, 2012; Yan *et al.*, 2013; 2015; 2014], we adopt the Schatten p-norm instead of the traditional trace norm to discover the low rank matrix. The Schatten p-norm ($0 < p < \inf$) of a matrix $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{S_p} = (Trace((\mathbf{M}'\mathbf{M})^{\frac{p}{2}}))^{\frac{1}{p}}$. If $p = 1$, it reduces to the trace norm. Eqn.(2) can be reformulated as:

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{V_a}, \mathbf{V_b}} & \|\mathbf{Y_a} - (\mathbf{U} + \mathbf{V_a})\mathbf{X_a}\|_F^2 + \|\mathbf{Y_b} - (\mathbf{U} + \mathbf{V_b})\mathbf{X_b}\|_F^2 \\ & + \lambda_1 \|\mathbf{U}\|_F^2 + \lambda_2 \|[\mathbf{V_a}; \, \mathbf{V_b}]\|_{S_p}^p \end{aligned} \tag{3}$$

[4] http://multiwordnet.fbk.eu/

Based on [Nie *et al.*, 2012], $\|\mathbf{V}\|_{S_p}^{p} = trace(\mathbf{V}'\mathbf{V}\mathbf{D})$, where $\mathbf{D}$ is the diagonal matrix $\mathbf{D} = \frac{2}{p}(\mathbf{V}'\mathbf{V})^{\frac{p-2}{2}}$. Then if we define $\mathbf{V}$ = [$\mathbf{V_a}$; $\mathbf{V_b}$] and $\mathbf{Y}$ = [$\mathbf{Y_a}$; $\mathbf{Y_b}$], Eqn.(2) becomes:

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{Y} - (\mathbf{U} + \mathbf{V})\mathbf{X}\|_F^2 + \lambda_1 \|\mathbf{U}\|_F^2 + \lambda_2 trace(\mathbf{V}'\mathbf{V}\mathbf{D}) \tag{4}$$

Optimization: Since Eqn.(4) involves the Schatten p-norm which is non-smooth and cannot be solved in a closed form, we adopt the alternating minimization algorithm to optimize the objective function with respect to $\mathbf{U}$ and $\mathbf{V}$ respectively in the following steps:

(I). Fix $\mathbf{V}$, optimize $\mathbf{U}$. By setting the derivative of Eqn.(4) *w.r.t.* $\mathbf{U}$ to zero, we have:

$$\frac{\partial}{\partial \mathbf{U}} = 2(\mathbf{U}\mathbf{X} + \mathbf{V}\mathbf{X} - \mathbf{Y})\mathbf{X}^T + 2\lambda_1 \mathbf{U}^T = 0$$

$$\mathbf{U} = (\mathbf{Y}\mathbf{X}^T - \mathbf{V}\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{X}^T + \lambda_1 \mathbf{I})^{-1}$$

(II). Fix $\mathbf{U}$, optimize $\mathbf{V}$. By setting the derivative of Eqn.(4) *w.r.t.* $\mathbf{V}$ to zero, we have:

$$\frac{\partial}{\partial \mathbf{V}} = 2(\mathbf{U}\mathbf{X} + \mathbf{V}\mathbf{X} - \mathbf{Y})\mathbf{X}^T + \lambda_2(\mathbf{V}\mathbf{D}^T + \mathbf{V}\mathbf{D}) = 0$$

Since $\mathbf{D}$ is a diagonal matrix, $\mathbf{D}^T = \mathbf{D}$, we have

$$\mathbf{V} = (\mathbf{Y}\mathbf{X}^T - \mathbf{U}\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{X}^T + \lambda_2 \mathbf{D})^{-1}$$

According to the optimization, we propose Algorithm 1 to solve the objective function of Eqn.(4).

Algorithm 1: An efficient iterative algorithm to solve the optimization problem in Eq.(4).

```
Input:
    The training data $\mathbf{X_a} \in \mathbb{R}^{d \times n_a}, \mathbf{X_b} \in \mathbb{R}^{d \times n_b}$;
    The training data labels $\mathbf{Y_a} \in \mathbb{R}^{m \times n_a}, \mathbf{Y_b} \in \mathbb{R}^{m \times n_b}$;
    Parameters $\lambda_1, \lambda_2, 0 \leq p \leq 2$.
Output:
    Optimized $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m \times d}$.
Set $t = 0$, randomly initialize $\mathbf{D}_t \in \mathbb{R}^{d \times d}$;
repeat
    Fix $\mathbf{V}$, Optimize $\mathbf{U}$:
    $\mathbf{U}_t = (\mathbf{Y}\mathbf{X}^T - \mathbf{V}_t\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{X}^T + \lambda_1\mathbf{I})^{-1}$;
    Fix $\mathbf{U}$, Optimize $\mathbf{V}$:
    $\mathbf{V}_t = (\mathbf{Y}\mathbf{X}^T - \mathbf{U}_t\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{X}^T + \lambda_2\mathbf{D}_t)^{-1}$;
    Update $\mathbf{D}_{t+1} = \frac{p}{2}(\mathbf{V}_{t+1}^T\mathbf{V}_{t+1})^{\frac{p-2}{2}}$;
    $t = t + 1$.
until Convergence;
```

6 Experiments

In this section we present the experimental results on the classification of the artworks with their corresponding textual information into positive and negative emotions.

6.1 Setup

For the late fusion with the weighted linear combination strategy, the weight a is tuned within the range $\{0.1, 0.2, 0.3, ..., 1\}$. For the joint flexible Schatten p-norm model, the parameters λ_1 and λ_2 are tuned within the range $\{10^{-3}, 10^{-2}, ..., 10^3\}$ and p is tuned within the range $\{0.01, 0.05, 0.1, 0.2, 0.5, 0.8, 1\}$. 5-fold cross-validations are used in all experiments and the best performance is reported.

6.2 Results

Since the metadata associated with a painting contains title and description, we perform three types of experiments: (i) visual + title; (ii) visual + descriptions; (iii) visual + title + descriptions. Table 1 shows the classification accuracy results using only title information. Since there are descriptions available only for 158 paintings, we consider this set to perform the experiments (ii) and (iii). Table 2 shows the classification accuracy results based only on descriptions information and Table 3 shows the classification accuracy results based on both paintings' title and description information. From the results, it can be observed that the combination of the textual and visual features improves the performance of the classification. Moreover, we observe that more than 3% improvement in accuracy is achieved by sharing patterns between visual and textual information through the joint flexible Schatten p-norm model compared with the late fusion strategy, which shows the effectiveness of our proposed method.

	Visual	Title	Late Fusion	Schatten p-norm
MART	0.753 ±0.043	0.582 ±0.044	0.755 ±0.043	0.783 ±0.024
deviantArt	0.763 ±0.038	0.674 ±0.038	0.776 ±0.038	0.801 ±0.028

Table 1: Classification accuracies considering only *title*. (All paintings are considered)

With these results, we investigated in which cases the title improves the emotional classification of abstract paintings. The textual features increase or even correct the classification results mostly in paintings displaying mixed visual cues. For example, light colors and smooth shapes generate a positive emotion, whereas dark colors and sharp edges generate a negative emotion [Sartori *et al.*, 2015]. The paintings with both light and dark colors, and/or smooth and sharp shapes are harder to classify. In such cases using titles and descriptions is expected to increase the classification accuracy.

	Visual	Description	Late Fusion	Schatten p-norm
MART	0.664 ±0.035	0.594 ±0.024	0.666 ±0.037	0.701 ±0.031
deviantArt	0.717 ±0.024	0.683 ±0.023	0.721 ±0.021	0.742 ±0.019

Table 2: Classification accuracies considering only *description*. (158 paintings with description are considered)

Figure 2 provides example of paintings together with their respective titles from each dataset. These paintings (MART dataset on the left, dA dataset on the right) are classified as positive when we use only the visual features. However, they are classified as negative when we use only the title information. When we apply the late fusion of visual and textual

	Visual	Title + Description	Late Fusion	Schatten p-norm
MART	0.664 ±0.035	0.474 ±0.033	0.664 ±0.023	0.697 ±0.022
deviantArt	0.717 ±0.024	0.658 ±0.031	0.725 ±0.033	0.751 ±0.021

Table 3: Classification accuracies considering *title* + *description*. (158 paintings with description are considered)

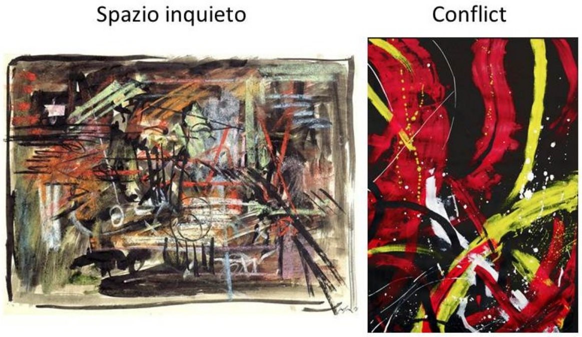
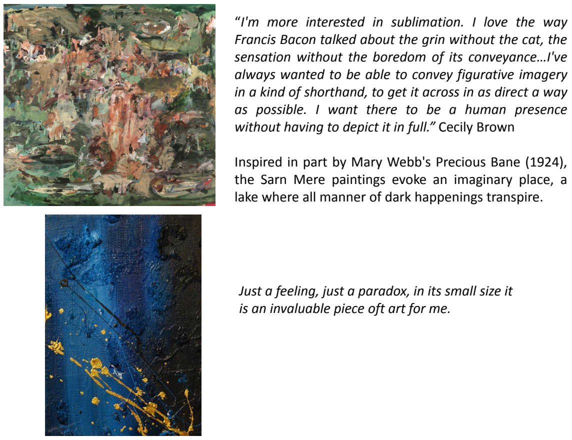


Figure 2: Two examples where title helped the classification algorithm. On the left: MART dataset, Spazio inquieto (Restless Space), 1953 - Emilio Vedova (*Courtesy of MART photographic archive, Rovereto.*) On the right: deviantArt dataset, Conflict, 2013 - Georgiana Beligan, phylactos.deviantart.com/ (*Courtesy of deviantArt.*)

features, the results are consistent with the ground truth, classifying these paintings as negative. We also observe that in both datasets, the paintings for which the descriptions helped the classification are mostly considered positive by people. Examples of such paintings from each dataset are provided in Figure 3. Both paintings, as they are composed of dark colours (dark green and dark blue) are considered as negative when using only visual features. However, when the descriptions are combined with the visual features the outcome is positive, which is consistent with human annotation.

By generally comparing both datasets, we observe that adding text as a feature is more effective for deviantArt than for MART. One reason may be that the descriptions in deviantArt are made by the artists themselves. Besides, only a few paintings are untitled in deviantArt (only 1%). On the contrary, in MART 10.4% of the paintings have no titles and the descriptions are usually made by others (e.g., art historians, art critics).

Finally, we also study the parameter sensitivity of the proposed method as demonstrated in Fig.4. Here, we fix $\lambda_1 = 1$ and analyze the regularization parameters λ_2 and p. As shown in Fig.4, the proposed method is more sensitive to p compared with λ_2, which confirms the importance of flexibility of our proposed Schatten p-norm model.



"I'm more interested in sublimation. I love the way Francis Bacon talked about the grin without the cat, the sensation without the boredom of its conveyance…I've always wanted to be able to convey figurative imagery in a kind of shorthand, to get it across in as direct a way as possible. I want there to be a human presence without having to depict it in full." Cecily Brown

Inspired in part by Mary Webb's Precious Bane (1924), the Sarn Mere paintings evoke an imaginary place, a lake where all manner of dark happenings transpire.

Just a feeling, just a paradox, in its small size it is an invaluable piece oft art for me.

Figure 3: Example where descriptions help on the classification. On the top: MART dataset, Study for Sarn Mere II, 2008 - Cecily Brown. (*Courtesy of MART photographic archive, Rovereto.*) On the bottom: deviantArt dataset, Rot, 2009 - Sias Est, excymiir.deviantart.com. (*Courtesy of deviantArt.*)

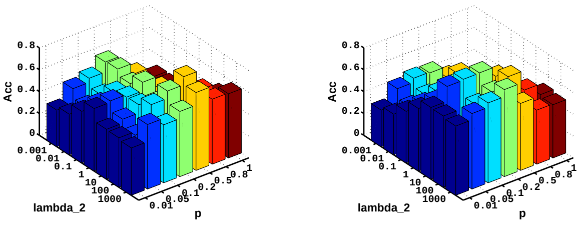


Figure 4: Sensitivity study of parameters on (left) MART dataset and (right) deviantArt dataset.

7 Conclusions

In this paper we showed that the metadata associated with abstract paintings is a useful addition to automatically identify the positive and negative emotions evoked by these paintings. We proposed a multimodal approach, in which we employed computer vision techniques and sentiment analysis to learn statistical patterns correlated to the valence emotions conveyed by abstract paintings. We proposed two approaches to combine visual and textual features. The first is late fusion based on weighted linear combination and the second is a novel joint flexible Schatten p-norm model which can exploit the sharing patterns between visual and textual information. We evaluated our method in two datasets of abstract artworks consisting of professional and amateur paintings respectively. We observed that using textual features (i.e., titles and artist descriptions) improves the accuracy in both approaches of combination.

A qualitative analysis showed that the combination of metadata associated with the paintings can help computers to efficiently recognize positive and negative emotions in the cases when the painting composition is mixed, for instance, when the painting is composed by dark colours (which conveys negative emotions) and light colours (which conveys positive emotions) at the same time.

Acknowledgments:

This work was partially supported by the MIUR Cluster project Active Ageing at Home, the EC project xLiMe and A*STAR Singapore under the Human-Centered Cyberphysical Systems (HCCS) grant.

References

[Artale *et al.*, 1997] A. Artale, B. Magnini, and C. Strapparava. Wordnet for Italian and its use for lexical discrimination. In *AI* IA 97: Advances in Artificial Intelligence*, pages 346–356. 1997.

[Baccianella and Sebastiani, 2010] A. E. S. Baccianella and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, 2010.

[Balahur *et al.*, 2011] A. Balahur, J. M. Hermida, A. Montoyo, and R. Munoz. Emotinet: a knowledge base for emotion detection in text built on the appraisal theories. In *NLDB*, pages 27–39, 2011.

[Franklin *et al.*, 1993] Margery B. Franklin, Robert C. Becklen, and Charlotte L. Doyle. The influence of titles on how paintings are seen. *Leonardo*, 26, No. 2:103–108, 1993.

[Gong *et al.*, 2014] Y. Gong, Q. Ke, M. L. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

[Guerini *et al.*, 2013] M. Guerini, L. Gatti, and M. Turchi. ”Sentiment Analysis: How to derive prior polarities from SentiWordNet”. In *EMNLP*, pages 1259–1269, 2013.

[Herbrich and Graepel, 2006] R. Herbrich and T. Graepel. Trueskill(TM): A bayesian skill rating system. *no. MSR-TR-2006-80*, 2006.

[Hristova *et al.*, 2011] E. Hristova, S. A. Georgieva, and M. Grinberg. Top-down influences on eye-movements during painting perception: the effect of task and titles. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces*, pages 104–115, 2011.

[Hwang and Grauman, 2012] S. J. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 100(2):134–153, 2012.

[Kandinsky, 1914] W. Kandinsky. *Concerning the Spiritual in Art.* Dover Books on Art History Series. Dover, 1914.

[Klein and Manning, 2003] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of ACL, 2003*, pages 423–430, 2003.

[Lang *et al.*, 1999] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. 1999.

[Leder *et al.*, 2006] H. Leder, C. Carbon, and A. Ripsas. Entitling art: Influence of title information on understanding and appreciation of paintings. *Acta psychologica*, 121(2):176–98, 2006.

[Liu *et al.*, 2011] N. Liu, E. Dellandrea, B. Tellez, and L. Chen. Associating textual features with visual ones to improve affective image classification. In *ACII*, 2011.

[Machajdik and Hanbury, 2010] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM Multimedia*, pages 83–92, 2010.

[Miller, 1995] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Millis, 2001] K. Millis. Making meaning brings pleasure: the influence of titles on aesthetic experiences. *Emotion*, 1(3):320, 2001.

[Moser, 2010] J. Moser. True skil library. https://github.com/moserware/Skills/, 2010.

[Nie *et al.*, 2012] F. Nie, H. Huang, and C. Ding. Low-rank matrix recovery via efficient schatten p-norm minimization. In *AAAI*, pages 655–661, 2012.

[Pianta *et al.*, 2008] E. Pianta, C. N. Girardi, and R. Zanoli. The textpro tool suite. In *LREC*, 2008.

[Sartori *et al.*, 2015] A. Sartori, V. Yanulevskaya, Akdag A. Salah, J. Uijlings, E. Bruni, and N. Sebe. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. *ACM TIIS*, 2015.

[Strapparava and Mihalcea, 2008] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *ACM Symposium on Applied Computing*, pages 1556–1560, 2008.

[The Museum of Modern Art, 2015] MoMA Highlights. The Museum of Modern Art, 2015. revised 2004, originally published 1999, p. 187.

[Yan *et al.*, 2013] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, 2013.

[Yan *et al.*, 2014] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe. Multi-task linear discriminant analysis for multi-view action recognition. *IEEE TIP*, 23(12):5599–5611, 2014.

[Yan *et al.*, 2015] Y. Yan, Y. Yang, D. Meng, G. Liu, W. Tong, A. G. Hauptmann, and N. Sebe. Event oriented dictionary learning for complex event detection. *IEEE TIP*, 24(6):1867–1878, 2015.

[Yanulevskaya *et al.*, 2008] V. Yanulevskaya, J. V. Gemert, K. Roth, A. Herbold, N. Sebe, and J. Geusebroek. Emotional valence categorization using holistic image features. In *IEEE ICIP*, pages 101–104, 2008.

[Yanulevskaya *et al.*, 2012] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe. In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. In *ACM Multimedia*, 2012.

[Zhao *et al.*, 2014] S. Zhao, Y. Gao, X. Jiang, H. Yao, T. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *ACM Multimedia*, pages 47–56, 2014.