Quantification of Cinematography Semiotics for Video-based Facial Emotion Recognition in the EmotiW 2015 Grand Challenge

Albert C. Cruz
acruz@cs.csubak.edu
Department of Computer and Electrical Engineering & Computer Science
California State University, Bakersfield
Bakersfield, CA 93311

ABSTRACT

The Emotion Recognition in the Wild challenge poses significant problems to state of the art auditory and visual affect quantification systems. To overcome the challenges, we investigate supplementary meta features based on film semiotics. Movie scenes are often presented and arranged in such a way as to amplify the emotion interpreted by the viewing audience. This technique is referred to as mise en scéne in the film industry and involves strict and intentional control of color palette, light source color, and arrangement of actors and objects in the scene. To this end, two algorithms for extracting mise en scéne information are proposed. Rule of thirds based motion history histograms detect motion along rule of thirds guidelines. Rule of thirds color layout descriptors compactly describe a scene at rule of thirds intersections. A comprehensive system is proposed that measures expression, emotion, vocalics, syntax, semantics, and film-based meta information. The proposed mise en scéne features have a higher classification rate and ROC area than LBP-TOP features on the validation set of the EmotiW 2015 challenge. The complete system improves classification performance over the baseline algorithm by 3.17% on the testing set.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications; I.4.m [**Image Processing and Computer Vision**]: Miscellaneous

Keywords

Mise en Scéne; syntax and semantics; EmotiW 2015 Challenge

1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ACM ICMI Proc. ACM Int'l. Conf. Multimodal Interaction
ACM 978-1-4503-3912-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2830592.

Emotion recognition is the study of computer algorithms that infer the underlying affect of a human from visual and auditory cues. We propose a comprehensive system for the Emotion Recognition in the Wild Challenge 2015 (EmotiW 2015) AFEW subchallenge [1] that combines vocalics, syntax, semantics, apparent expression, universal affect features, as well as features derived from cinematographic meta information. Publicly available data sets and grand challenges have energized the advancements to state of the art methods in the field of video-based affect recognition. Among the first such standard databases was the Japanese Female Facial Expression dataset (JAFFE) [2]. Later, it was realized that the expressions in most datasets were posed. Naturalistically expressed emotion must be captured so that a computer vision system can learn an accurate model to predict emotion in the wild [3]. The facial Emotion Recognition and Analysis challenge [4] was among the first such grand challenges to offer professionally acted data. Following this effort, the Audio Visual Emotion Challenge [5] and EmotiW [3] also offered more natural data. EmotiW is distinct from other challenges in that clips from AFEW are professionally acted scenes from movies. This is key to the novel methods presented in this approach.

Films use established semiotics to communicate a story to the viewing audience. As Frank Capra, a famous director, once said, "Film is one of three universal languages, the other two, mathematics and music." Clips in the data convey an emotion with more than just visual and auditory information. Nearly standard cues, referred to as *signs* in cinematography, supplement the ability of the audience to understand the affect of the actor [6]. Some techniques are so overused that they are cliché. We make significant contributions to the baseline approach including the incorporation of contextual information from object placement and lighting from cinematography semiotics, referred to as *mise en scéne* [6].

The rest of this work is organized as follows. Section 2 contains related work, motivation and contributions. Section 3 gives the technical approach and mise en scéne features are given in Section 3.1. Section 4 gives experimental parameters and results. Section 5 discusses the results. Section 6 concludes the paper and summarizes the key findings of the method.

2. RELATED WORK

In the following section, we summarize the top performers from previous EmotiW challenges.

EmotiW 2013 Challenge: M. Liu et al. [7] identified face regions with a PCA subspace approach, carried out subspace learning of appearance features with a Grassmannian manifold, and predicted labels with multiple one versus all partial less squares classifiers. S. Kahou et al. [8] applied different neural networks for each modality: a deep convolutional neural network for facial expression, a deep belief network for audio, a deep autoencoder for human actions, and a shallow network for mouth-specific features. K. Sikka et al. [9] combined SIFT, GIST and audio features, and carried out prediction with Multiple Kernel Learning and Support Vector Machine (SVM).

EmotiW 2014 Challenge: M. Liu et al. [10] computed histogram of oriented gradients, SIFT, and a convolution neural network. These features were projected into a Riemannian manifold before classification. J. Chen et al. [11] proposed histogram of orientated gradients of three orthogonal planes, and applied multiple kernel learning with SVM for classification. Sun et al. [12] extracted SIFT, LBP-TOP, pyramid of histograms of oriented gradients, and audio features from the openSMILE toolkit [13]. Prediction of affect was carried out at the decision level with a tiered re-voting scheme.

2.1 Motivation of Approach

Because the clips are from movies, actors in the scene will express their emotion in an overt way that can be understood by the audience. Despite this requirement, EmotiW 2015 still poses difficult challenges to state of the art facial emotion recognition methods. There are multiple characters in the same scene. There is no guarantee that the frame will focus on the person of interest who is expressing the emotion that needs to be predicted. The scene may abruptly shift from the subject's face. In at least one video, the majority of the clip is not focused on the face of the subject. There are greatly varying extrinsic imaging conditions that pose difficulties to face region of interest extraction algorithms. E.g., it may be too dark, the face region of interest may be too small or the face pose angle can be too great. All these factors are significant challenges to a video-only based affect recognition system.

Auditory cues provide discriminative information that can supplement an affect recognition system. However, multiple actors can speak in the same segment. For at least a few videos, the subject actor is silent and another actor is speaking. Additionally, since the clips are from movies, there may be a musical score and a glut of sound effects to immerse the viewing audience in the scene. These factors inhibit the ability of accurate quantification of auditory cues from the actor of focus.

Instead, we investigate the commonly used signs from the mise en scéne concept in cinematography and how they can supplement the auditory and visual affect projected in the scene. Consider a scene of an actor expressing fear, e.g. training set fear video # 010845040, a scene from Harry Potter (Figure 1-a). The actors eyebrows are raised in an expression of fear and there is an eerie voice. While this may be enough information to infer fear, there are a number of signs that contribute to this feeling. For instance, Harry Potter is gazing out of the window at something which is out of view of the audience–viewers can only guess what he is looking at and thinking about. The camera angle is spiraling toward his face indicating a loss of power-control affect. The tonal distribution of the image tends toward darkness and the color palette is limited to blues and greens. This is an example of how specific configurations of mise en scéne amplify the emotion being interpreted by the viewing audience. Proper execution of mise en scéne is entailed by strict and intentional control of lighting, costume, setting, color palette, props and the placement of objects and actors in the scene–referred to as *theatrical blocking* [6].

(a)

(b)

Figure 1: This figure should be viewed in color. Applications of key and color in cinematography. In general, the proportion of lighter tones are correlated with the valence conveyed in the scene. Projected affect is further controlled by the temperature of the light source. (a) Frame with letterboxing removed. A low keyed scene (dark with some light highlights) and cold tones from a high temperature light source convey fear and suspense. (b) A high keyed scene (very light with some dark highlights) and a low temperature light source used in a scene conveying happiness.

Previous work has explored encoding the emotional identity of movies from light source color [14]. In the Audio Visual Emotion Challenge 2014 [15], Kächele et al. outperformed the baseline approach focusing on meta information rather than features derived from audio or video. E.g., the ID of the subject, the length of the video, general motion and pixel differences, approximate weight, approximate age, and semantic content. In Dhall et al. [3], social event information, location and proximity of individuals in the scene were used for group affect analysis. To the best of our knowledge, we are the first to fuse cinematography-based meta information in a comprehensive system with vocalics, syntax, semantics, expression and emotion features.

2.2 Contribution

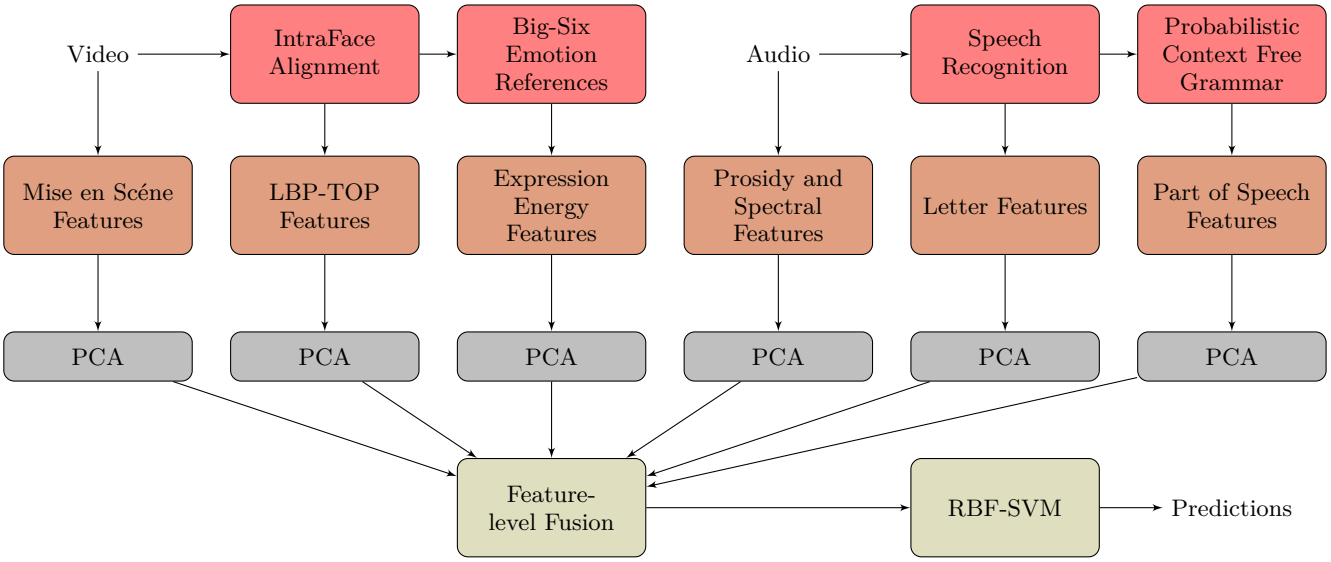


Figure 2: Overview of the approach. Features from left to right: mise en scéne configuration is leveraged as cinematographic meta information, apparent expression is extracted with LBP-TOP, emotion similarities are computed from references of faces expressing big six emotions, vocalics are extracted with prosidy and speech features, syntax is quantified with a letter histogram and semantics are quantified with a part of speech histogram. Fusion of modalities is carried out at the feature level and a SVM predicts emotion labels.

We are the first to incorporate mise en scéne based meta information in a facial emotion recognition system. Lighting, color and theatrical blocking features supplement vocalic, semantic, expression and emotion features. Two new features are proposed: (1) rule of thirds based motion history histograms that capture motion at rule of thirds guidelines and (2) rule of thirds based color layout descriptors that compactly represent the scene at rule of thirds intersections.

3. TECHNICAL APPROACH

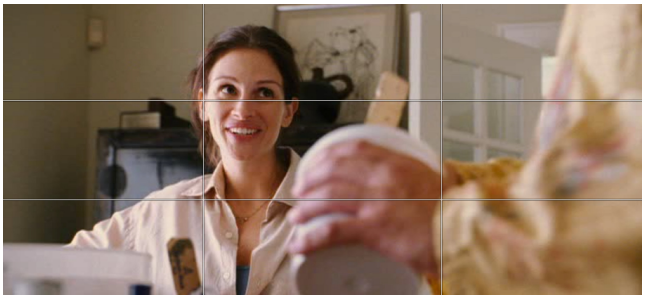
For mise en scéne features, key and overall color are measured with max-RGB and shades of gray color constancy algorithms. The proposed rule of thirds motion history histogram algorithm measures motion at the lines and intersections of rule of thirds guidelines. A color layout descriptor is computed at the rule of thirds guideline intersections to quantify the overall tone at these locations. For audio features, vocalics, syntax and semantics are extracted. openSMILE vocalic features are supplemented by additional features from the VoiceSauce toolbox [16]. The audio waveforms are parsed for syntax with Dragon Dictate for Mac, then further parsed for semantics by the Stanford Parser package [17]. For visual features, expression level and emotion level features are extracted. Apparent expression information is captured with LBP-TOP features. Similarities to big six reference emotions are computed with expression energy [18]. A system overview is given in Figure 2.

3.1 Mise en Scéne Features

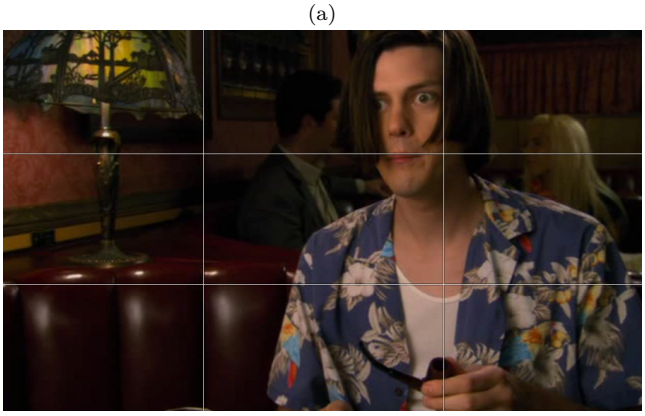
The video clips from the AFEW dataset are segments from professionally filmed movies that follow specific patterns expected by the audience. A general example of this are genre films, such as action, fantasy, mystery, etc., that follow a specific story pattern. Cinematographers also follow rules of thumb when filming scenes to augment the emotion being projected to the audience. It is not just conveyed through the expressions, vocalics and semantics of the actor of focus, but also how the scene is presented to the movie goer. The actors are expected to act and speak in a conspicuous or typical way based on the affective state of the character they are acting out, and so to must the properties of a scene be arranged to convey a specific emotion. E.g., if the director intends to convey fear, the scene will be poorly lit. If the cinematographer intends to convey happiness, the scene will be well lit and vibrantly colored. We consider two concepts of mood lighting: key and color.

Key is the proportion of light and dark in an image. High keyed scenes have a higher proportion of light tones and convey a feeling of warmth and energy. This is accomplished by having the entire scene well lit. E.g., the Wizard of Oz (1939) is full of vibrant colors. Low key lighting is the opposite, with a higher proportion of dark tones. Low key lighting conveys melodrama and suspense. Continuing with the Wizard of Oz example, low key lighting was used when the characters first met the evil antagonist of the story, the witch. These are serious scenes intended to convey apprehension. E.g., in Figure 1-a, low keyed lighting is used in a scene with fear.

When filming a movie, the director will use a specific color palette that fits the emotion being conveyed in the movie to augment the affect being expressed by the actors. Cinematographers generalize colors in terms of warm, such as red, yellow and orange, or cool, such as blue, green and violet. Scenes which are warm convey feelings of positive valence and positive attention. Scenes which are cool convey feelings of negative valence and negative predictability. When filming, this is achieved by using a color correction gel to achieve a specific light source temperature. E.g., in Figure 1-b, a low temperature light source is used.



(a)



(b)

Figure 3: Applications of rule of thirds in cinematography. In a scene, points of interest are generally placed near two sets of equally spaced lines, with the sets perpendicular to each other. (a) Frame with letterboxing removed. The face of the actor is placed near the intersection in the upper left-hand corner. (b) The actor's body is placed along the right line.

Another cinematographic technique considered in our approach is theatrical blocking, specifically rule of thirds. This concept asserts that an image should be divided into nine equally sized sub-regions by two equally spaced vertical lines and two equally spaced horizontal lines. Points of interest in the scene are placed along the lines or intersections of the lines. E.g. the face of the actor is often placed at the intersection of the guide lines, see Figure 3. A summary of the features used in the method follows:

- Normalized color histogram of the entire video, summed across all frames
- Light source tone computed with max-RGB, averaged across all frames
- Light source tone computed with shades of gray, averaged across all frames
- Motion history histogram weighted by distance to rule of thirds guidelines
- Color layout descriptor computed at each rule of thirds intersection

3.1.1 Color Palette and Key Features

The first feature extracted is a color histogram with each channel quantized to sixteen bins. This gives overall key and color of the frame. However, in filmography, the color palette is managed by color temperature of light sources, motivating more in depth representation of the colors in the frame. This information is not provided with the data and must be estimated. Color constancy is the study of object tones measured independently of light source color. Methods attempt to estimate the light source tone and use this information to correct the tones in the image. We apply the methods which compute the light source color $\mathbf{l}$. Assuming that the average light reflectance in the scene is achromatic:

$$k \int_{\lambda} l(\lambda)\, \mathbf{c}(\lambda)\, d\lambda = k\mathbf{l} \tag{1}$$

where λ is the wavelength, k is a constant based on the incident light and $\mathbf{c}$ is the camera sensitivity.

With the max-RGB color constancy method [19] this is estimated as:

$$k\mathbf{l} = \max_{\mathbf{p}} \mathbf{f}(\mathbf{p}) = \left| \max_{\mathbf{p}} f_r(\mathbf{p}), \max_{\mathbf{p}} f_g(\mathbf{p}), \max_{\mathbf{p}} f_b(\mathbf{p}) \right| \tag{2}$$

where f_r, f_g, and f_b are the red, green and blue channels of the image $\mathbf{f}$ respectively. Each channel is separately searched for the lightest tone, and the maximum value in each of these channels is taken to be the respective light source color component. The second method employed is called shades of gray. It computes the light source color as a function of the L_p norm of the tone:

$$k\mathbf{l} = \left(\frac{1}{N_{\mathbf{p}}} \sum_{\mathbf{p}} \mathbf{f}(\mathbf{p})^p \right)^{1/p} \tag{3}$$

where p is the value of the L_p norm and $N_{\mathbf{p}}$ is the number of pixels. For $p = \infty$, Equation 3 is equal to Equation 2 [20]. The results of these two equations are computed on a frame by frame basis.

3.1.2 Rule of Thirds Based Features

Theatrical blocking is measured by extracting features of the scene at the lines and intersections of the rule of thirds guidelines. Objects that the viewing audience should focus on are placed at the lines and intersections of these guidelines. When assuming this model, regions far from the guidelines are background and not a part of the regions of interest.

The first feature proposed quantifies activity of the actor by measuring motion along the rule of thirds guidelines. Frequently used approaches involve quantifying motion with optical flow or with a feature-descriptor matching process. However, these methods can be computationally undesirable given the large resolution of the frames. Instead, we use motion history images and histograms because of their ability to succinctly represent spatiotemporal information. Among the earliest uses of motion history images was [21], where motion history images (MHI) were used to describe action scenes as a single image representation. Variations of the motion image concept continue to be used for emotion recognition (MHH in depression recognition [22]). The motion energy image is computed as:

$$m_t(\mathbf{p}) = \begin{cases} 1 & |\mathbf{f}_t(\mathbf{p}) - \mathbf{f}_{t-1}(\mathbf{p})| > \tau \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

where m_t is the motion history image at time t. $\mathbf{f}_t$ is the frame color at time t. We modify the MHI concept to focus

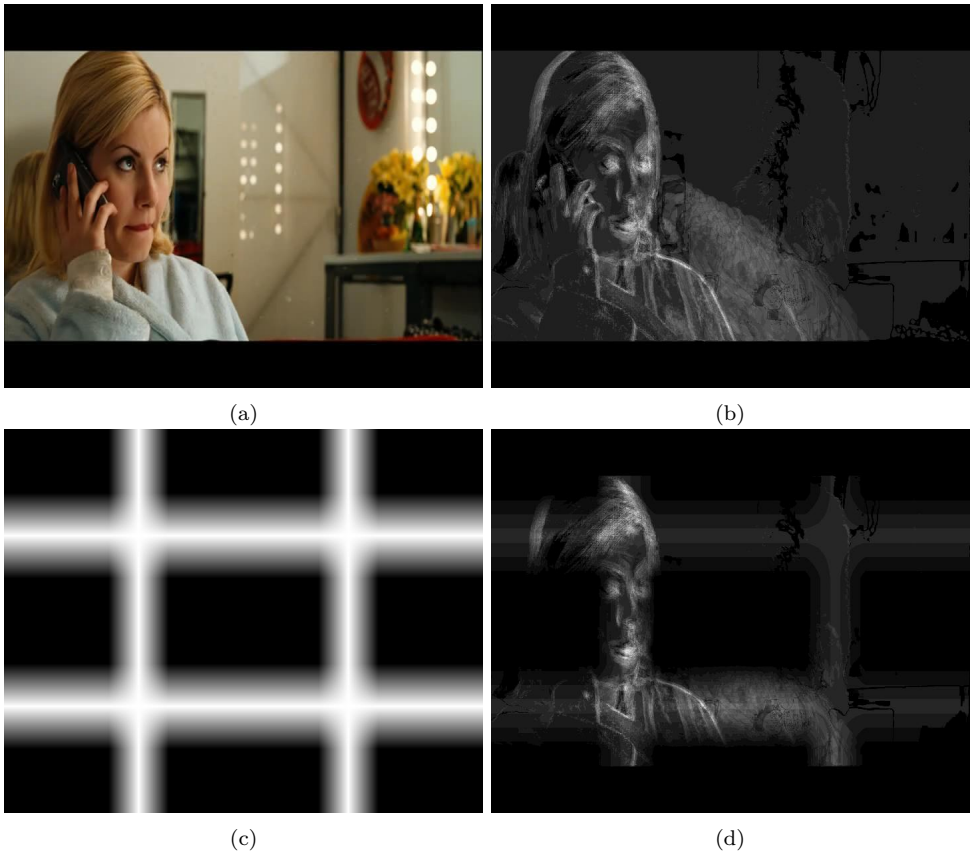


Figure 4: Rule of thirds motion history histogram. (a) Original frame. (b) Summation of motion history images. (c) Weighting scheme based on rule of thirds guidelines. (d) Result of weighting (b) with (c).

strictly on temporal changes near the rule of thirds guidelines (spatial information is handled with the CLD descriptor below). The time dimension in Equation 4 is marginalized, and weighted:

$$\tilde{m}(\mathbf{p}) = \sum_{k} g(\mathbf{p})\, m_k(\mathbf{p}) \tag{5}$$

where $g(\mathbf{p})$ is an image containing the weight values. A pixel is weighted inversely proportional to its distance to one of the rule of thirds guidelines (see Figure 4-c). As the final step, a histogram is taken of $\tilde{m}$. The proposed method is significantly different from previous MHI and MHH algorithms. The output of this method is a motion magnitudes histogram capturing temporal motion of theatrical blocking. Previous MHI algorithms captured spatiotemporal motion represented as an image. An example of the method is given in Figure 4. Letterboxing is removed before computing the locations of the guidelines. Note that this feature only measures motion. The following feature measures color palette.

Spatial information is quantified at the rule of thirds intersections with color layout descriptors. Color layout descriptors were defined with the MPEG-7 standard as a method for quantification of tone of a video at a glance [23]. First, an image is reduced to 8×8 resolution for each channel by averaging the tone of the regions before subsampling. Then, the image is converted to YCbCr color channels. The DCT of each channel is taken, and a zig-zag traverse re-orders components by frequency. Typically, some high frequency components are removed. In this work, we do not reduce the amount of information at this stage; PCA is used to compute the most variant axes later on. It should remove high frequency components as most energy is located in the lower frequency spectrum. This feature captures theatrical blocking as well as color palette information and keying.

3.2 Audio Features

3.2.1 Vocalic Features

We supplement the baseline audio features from openSMILE with features from the VoiceSauce toolbox [16]. Features are computed for the entire video then averaged to produce a single feature vector. A summary of these features are as follows:

- All baseline audio features
- Fundamental frequency, fundamental frequency frequencies and bandwidths of the first four formants
- Amplitudes of harmonics
- Root mean square of waveform at every frame over a variable window equal to five pitch pulses
- Ceptral peak prominence
- Subharmonic-to-harmonic ratio

Table 1: Results on EmotiW 2015 for varying modalities, fusion results and results on the testing fold. CR: classification rate. FPR: false positive rate. Bold: best performer for set. Underline: second best performer for set.

Validation Set						
	CR	FPR	Precision	Recall	F-measure	ROC Area
Vocalics (openSMILE, VoiceSauce)	0.272	0.140	0.274	0.272	0.235	0.530
Syntax + Semantics (Dragon, Stanford Parser)	0.232	0.151	**0.350**	0.232	0.134	0.515
Expression (LBP-TOP)	0.315	**0.129**	0.266	<u>0.315</u>	<u>0.267</u>	0.528
Emotion (Expression Energy)	0.237	0.152	0.173	0.237	0.174	0.526
Meta information (Mise en Scéne)	<u>0.324</u>	0.152	0.231	0.286	0.227	<u>0.587</u>
Feature-level Fusion	**0.361**	<u>0.130</u>	<u>0.300</u>	**0.335**	**0.301**	**0.668**
Testing Set						
	CR	FPR	Precision	Recall	F-measure	ROC Area
Feature-level Fusion	0.425	0.103	0.625	0.203	0.282	-

3.2.2 Syntax and Semantic Features

The focus of facial affect recognition is the prediction of emotion from non-verbal cues. However, a significant part of communication is delivered via language. Thus, we supplement the approach with a system that computes the syntactic and semantic meaning of speech uttered in the video. As stated previously, because the video clips are from movies, the actors can be expected to act and speak in a conspicuous or typical way. E.g., if a person is angry they will curse. To compute these features, syntax is extracted from vocalics and then the syntax is processed into a parse tree to find parts of speech. Both the syntactic and semantic features are used for the approach.

Dragon commercial software was successfully employed in the Audio Visual Emotion Challenge 2014 for speech recognition [15]. In this work, we use Dragon Dictate for Mac to compute the words from the video clips. At this stage in syntax and semantic processing, a letter histogram is computed as a feature. The sentences from the commercial software are parsed into a tree with an unlexicalized probabilistic context-free grammar method [17]. A histogram counting the frequency of parts of speech is used as a feature. A summary of the features in this component of the algorithm are as follows:

- Unnormalized letter histogram with non-alpha character omitted
- Unnormalized histogram of specific parts of speech: coordinating conjunction, cardinal number, determiner, existential there, preposition or subordinating conjunction, adjective, comparative adjective, superlative adjective, model, singular noun, plural noun, singular proper noun, plural proper noun, predeterminer, progressive ending, personal pronoun, possessive pronoun, adverb, comparative adverb, superlative adverb, particle, symbol, interjection, base form verb , past tense verb, present participle verb, past participle verb, non-3rd person singular present verb, 3rd person singular present verb, Wh-determiner, Wh-pronoun, possessive wh-pronoun, and wh-adverb
- Unnormalized histogram of general parts of speech: adjectives, nouns, verbs, and adverbs

3.3 Video Features

Low-level facial information is extracted with baseline LBP-TOP features. The reader is referred to [24] for an in-depth explanation of the algorithm. These features are supplemented with higher-level facial emotion information that computes each frame's similarity to one of the six basic emotions. The similarity is computed to a reference face that is generated from training. The reference face represents the typical expression of one of the big six emotions. A summary of the visual features used in the approach are as follows:

- LBP-TOP features from baseline
- Six similarity scores to big six emotion references

3.3.1 Emotion Features

In this section, we discuss the expression energy method that computes the similarity to one of the big six emotions. This method was first proposed in [25] and was shown to have promising generalization performance in a person-independent setting. The face image $\mathbf{f}$ is compared to reference faces of six-basic emotions. Because the six-basic emotions are universal, comparisons to representations of the six-basic emotions may be a more reliable encoding of affective cues than apparent facial information.

Let Φ be the set of references: $\Phi = \{\text{happy, sad, fearful, angry, surprised, disgusted}\}$. Let e be one of the emotions in Φ. Let r_e be the reference of a characteristically expressed emotion e. In [25], $\mathbf{f}$ was binary classified as either a given reference e or not: $\Phi - \{e\}$. This was carried out with a SVM, and the decision value was used as a feature for r_e. This was carried out once for each emotion in Φ. In our method, we compute the expression energy of $\mathbf{f}$ directly to each emotion in Φ and take the scores to be the feature vector. The score is a comparison of $\mathbf{f}$ to one of the emotions in Φ:

$$x_e = d\left(\mathbf{f}, r_e\right) \tag{6}$$

where e is the emotion.

The avatar reference image concept [26] is used to create a reference of a face from the training data. Avatar reference images originally created an expressionless reference face for registration purposes. Instead, we generate a reference of each r_e that positively expresses e.

Let $\mathbf{f}_e$ be an image expressing e. To estimate r_e we take the mean:

$$r_e\left(\mathbf{p}\right) = \frac{1}{N}\sum_{\mathcal{D}_e}\mathbf{f}_e\left(\mathbf{p}\right) \tag{7}$$

where $\mathcal{D}_e$ is the set of face regions of interest images positively expressing e, and N is the number of images in $\mathcal{D}_e$.

Then, SIFT-Flow [27] aligns each face in $\mathcal{D}_e$ to r_e. The process iterates by rewarping $\mathcal{D}_e$ to the avatar reference image.

We calculate d in Equation 6 as the energy of warping $\mathbf{f}$ to r_e with a dense warp of SIFT features. Let $\mathbf{w}$ be the flow field calculated with a modified SIFT Flow objective function:

$$E(\mathbf{w}(\mathbf{p})) = \sum_{\mathbf{p}} \|\mathbf{s}_{r_e}(\mathbf{p}) - \mathbf{s}_{\mathbf{f}}(\mathbf{p} + \mathbf{w}(\mathbf{p}))\|_1 \quad (8)$$
$$+\alpha \sum_{\mathbf{p}} \left((u(\mathbf{p}) - \mu_x)^2 + v(\mathbf{p} - \mu_y)^2 \right)$$

where E is the objective function that is minimized. α is a parameter. $\mathbf{s}_{\mathbf{f}}(\mathbf{p} + \mathbf{w}(\mathbf{p}))$ is the dense SIFT feature of $\mathbf{f}$ at $\mathbf{p}$ offset by $\mathbf{w}(\mathbf{p})$. μ_x and μ_y are the mean motion in the horizontal and vertical directions of $\mathbf{w}$ respectively. After optimal warp, Equation 8 is taken to be d.

The first term reduces the error of r_e and a candidate SIFT feature in $\mathbf{f}$. If the SIFT features are equal, the first term is zero. Non-zero values capture the changes from r_e to $\mathbf{f}$. The second term constrains the flow field by minimizing the magnitude of all of the flow vectors. We subtract the mean from the magnitude to account for slight translation errors from region of interest detection. Otherwise, translation errors would falsely approximate dissimilarity proportional to the size of the image. Because we are subtracting the mean, the effects of whole plane translation errors are reduced, and a fine registration process is not required. Non-zero values of Equation 8 approximate dissimilarity to reference r_e.

4. RESULTS

For mise en scéne features, τ in Equation 4 is 13. For avatar reference image, we iterated the algorithm three times because it was the optimal parameter found in [26]. RBF SVM was implemented with LibSVM [28]. For the SVM, the cost c was 3.5481 and γ was 0.3162. These values were selected with a grid search on the validation set. For rule of thirds color layout descriptor (CLD), CLD is computed in 100×100 sub-regions of the image placed at the rule of thirds intersections.

Results are given in Table 4 for varying modalities on the validation set, feature-level fusion on the validation set, and final results on the testing set. ROC area is not available for the testing fold results because the labels are not public. Expression (LBP-TOP) features are the same features provided in the baseline, except the classifier stage involves PCA dimensionality reduction and an RBF-SVM for classification. In depth results with classification rate, false positive rate, precision, recall, F-measure and ROC area give a full understanding of performance for each modality. Fusion of all modalities and mise en scéne features give the best classification rate. Fusion and expression give the least number of false positives. Semantics and fusion give the best precision. Expression and fusion give the best recall, F-measure. Fusion and mise en scéne features give the best ROC area. A confusion matrix of the testing results is given in Table 2.

5. DISCUSSION

The syntax and semantic features were the worst performers. There is no mechanism to isolate the significant speaker in the scene, so vocalic features measured background noise and music. In line with the mise en scéne concept, musical score key should predict apparent affect. However, a number of the videos did not have music. Additionally, some background music had lyrics which were parsed by the semantics modality.

The proposed mise en scéne features outperformed the baseline LBP-TOP features when comparing the modalities without fusion. But, there were still a significant number of misclassifications. Expressionist films are heavily stylized, with extremely typical and discriminative configurations of mise en scéne. However, a movie does not have to follow expressionist style. The scene can be presented as-is with no distortions to light, color, etc. This style of cinematography is referred to as realism. Additionally, artistic value can be generated from intentionally violating cliché mise en scéne patterns. Movie style can also vary from director to director. Meta information estimating the cinematography style or director identity could further supplement mise en scéne features for movie-based affect recognition.

We further investigate the impact of mise en scéne features with a Q-statistic [29]:

$$q_{i,j} = \frac{(\tilde{q}_{00}\tilde{q}_{11} - \tilde{q}_{10}\tilde{q}_{01})}{(\tilde{q}_{00}\tilde{q}_{11} + \tilde{q}_{10}\tilde{q}_{01})} \quad (9)$$

The Q-statistic compares the success and failure rates of two different single modality feature methods, i and j. $\tilde{q}_{00}$ is the number of videos where both algorithms failed. $\tilde{q}_{11}$ is the number of videos where both algorithms succeeded. $\tilde{q}_{01}$ is the number of videos where i failed and j succeeded, and vise versa. $q_{i,j}$ is valued $[-1, 1]$. Values approaching -1 indicate that both i and j made identical decisions. Values approaching 1 indicate that i and j made opposite decisions, and should compliment each other if fused. Q-statistic results are given in Table 3. There is promise for pairing of mise en scéne with semantics, syntax, expression and emotion modalities. There are no negative values.

6. CONCLUSION

Table 2: Confusion matrix of results from the EmotiW 2015 challenge. Ang.: angry. Dis.: disgust. Hap.: happy. Neut.: neutral. Sur.: surprise.

	Ang.	Dis.	Fear	Hap.	Neut.	Sad	Sur.
Ang.	45	1	5	15	6	4	3
Dis.	4	0	2	7	13	3	0
Fear	9	0	33	10	7	6	1
Hap.	15	2	3	46	33	8	1
Neut.	6	1	14	31	85	21	1
Sad	14	1	6	9	20	19	2
Sur.	1	0	3	8	9	5	1

Table 3: Q-statistic values for each modality. Positive values of $q_{i,j}$ approaching 1 indicate disparate classifications, showing promise for fusion. i = Mise en Scéne.

Modality j	$\tilde{q}_{00}$	$\tilde{q}_{01}$	$\tilde{q}_{01}$	$\tilde{q}_{11}$	$q_{i,j}$
Vocalics	234	73	48	16	0.033
Syntax + Semantics	259	48	50	14	0.203
Expression	222	85	37	27	0.312
Emotion	264	43	44	20	0.472

We proposed a multimodal system that measured expression, emotion, vocalics, syntax, semantics, and film-based meta information, referred to as mise en scéne features. Two algorithms for extracting mise en scéne information were proposed. Rule of thirds based motion history histograms detected motion along rule of thirds guidelines. Rule of thirds color layout descriptors compactly described a scene at rule of thirds intersections. The system improved classification performance over the baseline algorithm by 3.17% on the testing set.

7. REFERENCES

[1] A. Dhall, O. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *ACM ICMI Workshops*, 2015.

[2] M. J. Lyons, "Automatic classification of single facial images," *IEEE Trans. PAMI*, vol. 21, no. 12, pp. 1357–1362, 1999.

[3] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[4] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. SMC, Part B: Cybernetics*, vol. 42, no. 4, pp. 966–979, 2012.

[5] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 - The first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction Workshops*, vol. 6975, pp. 415–424, 2011.

[6] R. Edgar-Hunt, J. Marland, and S. Rawle, *The Language of Film*. Fairchild Books, 2005.

[7] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen, "Partial least squares regression on grassmannian manifold for emotion recognition," in *ACM ICMI Workshops*, pp. 525–530, 2013.

[8] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, R. Memisevic, P. Vincent, A. Courville, and Y. Bengio, "Combining modality specific deep neural networks for emotion recognition in video," in *ACM ICMI Workshops*, pp. 543–550, 2013.

[9] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *ACM ICMI Workshops*, pp. 517–524, 2013.

[10] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild," in *ACM ICMI Workshops*, pp. 494–501, 2014.

[11] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *ACM ICMI Workshops*, pp. 508–513, 2014.

[12] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu, "Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild," in *ACM ICMI Workshops*, pp. 481–486, 2014.

[13] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM Multimedia*, pp. 835–838, 2013.

[14] L. Canini, S. Benini, P. Migliorati, and R. Leonardi, "Emotional identity of movies," in *IEEE Int'l. Conf. Image Processing*, pp. 1821–1824, 2009.

[15] M. Kächele and M. Schels, "Inferring depression and affect from application dependent meta knowledge," in *ACM Multimedia Workshops*, pp. 41–48, 2014.

[16] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," in *Int'l. Congress of Phonetic Sciences*, pp. 1846–1849, 2011.

[17] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks," in *Conf. Empirical Methods on Natural Language Processing*, 2014.

[18] A. C. Cruz, B. Bhanu, and N. Thakoor, "Facial emotion recognition with expression energy," in *ACM ICMI Workshops*, pp. 457–464, 2012.

[19] E. H. Land, "The retinex theory of color vision," in *Edwin H. Land's Essays* (M. McCann, ed.), pp. 53–60, Society for Imaging Science & Technology, 1993.

[20] J. Weijer, T. Gevers, and A. Gijsenij, "Edge-Based Color Constancy," *IEEE Trans. on Image Processing*, vol. 16, no. 9, pp. 2207–2214, 2007.

[21] J. W. Davis, "Recognizing Movement using Motion Histograms," Tech. Rep. 487, MIT, 1998.

[22] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression," in *ACM ICMI Workshops*, pp. 21–30, 2013.

[23] E. Kasutani and A. Yamada, "The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval," in *IEEE Int'l. Conf. Image Processing*, vol. 1, pp. 674–677 vol.1, 2001.

[24] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. PAMI*, vol. 29, no. 6, pp. 1–14, 2007.

[25] A. Cruz, B. Bhanu, and N. Thakoor, "One shot emotion scores for facial emotion recognition," in *IEEE Int'l. Conf. Image Processing*, pp. 1376–1380, Oct 2014.

[26] S. Yang and B. Bhanu, "Understanding discrete facial expressions in video using an emotion avatar image," *IEEE Trans. SMC, Part B: Cybernetics*, vol. 42, no. 4, pp. 920–992, 2011.

[27] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: dense correspondence across scenes and its applications," *IEEE Trans. PAMI*, vol. 33, no. 5, pp. 978–994, 2011.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[29] X. Zhou and B. Bhanu, "Integrating face and gait for human recognition at a distance in video," *IEEE Trans. SMC, Part B: Cybernetics*, vol. 37, no. 5, pp. 1119–1137, 2007.