Identification and Extraction of Digital Forensic Evidence from Multimedia Data Sources Using Multi-Algorithmic Fusion

Shahlaa Mashhadani[1, 2], Nathan Clarke[1, 3] and F. Li[4]

[1] Centre for Security, Communications and Network Research, Plymouth University, Plymouth, UK
[2] Computer Science Department/Collage of Education for Pure Science/ Ibn Al Haytham/ Baghdad, Iraq.
[3] Security Research Institute/Edith Cowan University/Perth, Western Australia, Australia.
[4] School of Computing, University of Portsmouth, Portsmouth, UK
shahlaa.mashhadani @ plymouth.ac.uk, N.Clarke @ plymouth.ac.uk, fudong.li@port.ac.uk

Keywords: Digital forensics, Multimedia Forensics, Automatic Image Annotation, Fusion.

Abstract: With the enormous increase in the use and volume of photographs and videos, multimedia-based digital evidence has come to play an increasingly fundamental role in criminal investigations. However, given the increase in the volume of multimedia data, it is becoming time-consuming and costly for investigators to analyse the images manually. Therefore, a need exists for image analysis and retrieval techniques that are able to process, analyse and retrieve images efficiently and effectively. Outside of forensics, image annotation systems have become increasingly popular for a variety of purposes and major software/IT companies, such as Amazon, Microsoft and Google all have cloud-based image annotation systems. The paper presents a series of experiments that evaluate commercial annotation systems to determine their accuracy and ability to comprehensively annotate images within a forensic image analysis context (rather than simply single object imagery, which is typically the case). The paper further proposes and demonstrates the value of utilizing a multi-algorithmic approach via fusion to achieve the best results. The results of these experiments show that by existing systems the highest Average Recall was achieved by imagga with 53%, whilst the proposed multi-algorithmic system achieved 77% across the selected datasets. These results demonstrate the benefit of using a multi-algorithmic approach.

1 INTRODUCTION

Digital images are now considered as a significant feature of many security systems, playing a major role in the forensic investigation of crimes (Redi et al. 2011). In the U.K., in addition to private security, there are now almost six million closed-circuit television (CCTV) systems covering public places including 750,000 in 'sensitive locations' such as banks, police stations, office buildings, and prisons, and public places such as airports, shopping centers, restaurants, and traffic intersections. This produces a vast volume of images photographic and video-based content (Forensicsciencesimplified.org 2016) and (Singh 2015). In addition, one trillion photos were taken in 2015 (Worthington 2015). This significant increase in the number of images have occurred, because of the increase of storage media, in addition to the cost of capturing pictures has become free. Consequently, massive digital images of evidence or crime scenes have to be investigated.

Within criminal investigations, such evidence can be vital in information gathering and in determining innocence or guilt. However, with such a volume of data to analyse, it can often be highly time-consuming. Understanding and interpreting such imagery can also place a huge burden upon the investigator. Whilst many forensic tools exist, such as EnCase, FTK, P2 Commander, Autopsy, HELIX3, and Free Hex Editor, their focus to date has been upon string-based examination, with image-based analysis restricted to optical character recognition and explicit image detection (Al Fahdi et al. 2016). Consequently, an investigator needs a more efficient and effective capability to interpret, analyse, and retrieve images from large repositories in an accurate and timely manner in order to solve criminal cases such as child abduction, stealing a customer's money bag in a bank, car theft and etc.

There are two main methods for retrieving images: retrieval by image content (image example) and retrieval by words (annotations). The former is referred to as Content-Based Image Retrieval (CBIR) and the latter, Annotation Based Image Retrieval (ABIR). CBIR is suitable for retrieved images such as X-ray pictures and faces of criminals from a video that record in a crime scene. However, there are two main shortcomings of CBIR. The first is CBIR cannot deal with applications that contain more semantic relationships even after adopting comprehensive image processing techniques. For instance, to retrieve images that related to the “Iraq war”, it is difficult to determine the kind of query image that can give acceptable and precise results. This is because the concepts cannot be fully represented by visual features. The other shortcoming is the CBIR premise; an example image must be available for the user, while in ABIR a user can simply compose queries using natural language (Inoue 2004). ABIR can itself be divided into two parts, Automatic Image Annotation (AIA) and query processing (Hidajat 2015). The main objective of the AIA is to determine the best annotations that can be used to describe the visual content of an untagged or wrongly tagged image (Tian 2015).

The ability for an investigator to search based upon keywords (an approach that already exists within forensic tools for character-based evidence) provides a simple and effective approach to identify relevant imagery. However, the focus of previous work in AIA has been focussed upon the general domain of image analysis, rather than focusing on the specific requirements that exist in a forensic image analysis context. Within the general context, there are a number of commercial AIA systems such as Google Cloud API (Google Cloud Platform 2017) and Clarifai (Calrifai 2018).

The aim of this paper is two-fold. To understand and evaluate the performance of current commercial AIA systems and secondly, to determine whether a multi-algorithmic approach to classification would improve the underlying performance. The reasons for using the multi-algorithmic approach are to increase annotation accuracy, improve the retrieval performance and collect different annotations for the same image (synonyms for the same object such as car and vehicle).

The remainder of the paper is organized as follows. Section 2 provides an overview of the current state-of-the-art within an AIA. Building upon this, Section 3 presents the research hypothesis and methodology. Section 4 presents the experimental results, with Section 5 providing a discussion of the approach and areas for future development. The conclusion is presented in Section 6.

2 LITERATURE REVIEW

The authors were unable to identify any studies that have focused upon AIA for the specific purpose of forensic image analysis. Only (Lee et al. 2011) deals with a particular forensic image database containing a large collection of tattoo images (64,000 tattoo images, provided by the Michigan State Police). They achieved 90.5% retrieval accuracy; however, the retrieval performance was affected by low-quality query images, such as images with low contrast, uneven illumination, small tattoo size, or heavy body hair covering the tattoo. To overcome the low quality of such images. They employed image annotation to improve the results; however, they depended on manual image annotation, which is time-consuming and is deemed unsuitable when dealing with a large volume of images. The performance of AIA systems is measured in two ways: annotation validation and retrieval performance. Annotation validation is measured by equation 1.

$$\text{Precision } = \frac{\text{number of correct words}}{\text{number of annotation words}} \qquad (1)$$

Whereas, retrieval performance is measured in terms of three parameters: precision (P), recall (R) and F-measure (F), as defined in equations 2, 3, and 4, respectively.

$$\text{Precision} = \frac{\text{Number of relevent images retreived}}{\text{Total number of images retreived}} \qquad (2)$$

$$\text{Recall} = \frac{\text{Number of relevent images retreived}}{\text{Total relevant images in collection}} \qquad (3)$$

$$\text{F} = 2 * \frac{\text{precision}*\text{recall}}{\text{presion}+\text{recall}} \qquad (4)$$

In recent years, several studies have focused on the AIA as illustrated in Table 1. The studies utilized a number of different datasets with differing compositions, making it difficult to compare their performances directly. It does, however, provide an understanding of the general performance that can be achieved. With respect to the dataset, several authors examined their systems using the Corel 5k dataset (Li et al. 2012) (Xie et al. 2013) (Zhang et al. 2013) (Bahrami and Abadeh 2014) (Zhang 2014b) (Zhang 2014a) (Hou and Wang 2014) (Yuan-Yuan et al. 2014) (Murthy et al. 2014) and (Tian 2014). The study (Hou and Wang 2014) achieved 80% P, which is higher than the results of other studies using the same dataset with a single or double classifier(s). This can be explained by the fact that multiple classifiers can improve accuracy results by combining the advantages of all implemented classifiers. In addition, the use of multiple classifiers affords the chance to generate different results that can be fused together in order to achieve high accuracy of annotation results. (Bahrami and Abadeh 2014) (Zhang, 2014a) and (Tian, 2014) used the same dataset (Corel 5k) and segmentation method (the normalized cut algorithm) and their P were 34%, 25%, and 24% respectively. These varying results can be attributed to using different types of classifiers and variation in feature extraction methods. The research studies by (Zhang, 2014b) and (Zhang, 2014a) applied the same segmentation approach, feature extraction methods and dataset (Corel 5K) the former study reported 34% P and 24% R using linear regression for the classification task. The latter utilized non-linear regression and the accuracies were varied by implementing the Gaussian kernel and the polynomial kernel functions.

Table 1: Summary of AIA Studies.

Authors	**Segmentation Method**	**Feature Extraction**	**Classifier Name**	**Performance (%)**			**Dataset Name**	**Images No.**
				P	**R**	**F**		
Hidajat 2015	Gaussian Mixture model	SIFT	SVM	88	65	76	LAMDA	541
Sumathi and Hemalatha 2011	-	JEC feature extraction	SVMs	77	35	51	Flicker	500
Li et al. 2012	Dividing image into blocks (16*16)	Color: 24 color features Texture: 12 texture features	Hybrid Generative/Discriminative Model	32	28	-	Corel	5000
Xie et al. 2013	-	12 visual features	Two-phase generation model (LIBSVM, co-occurrence measures)	34 44	51 50	41 47	Corel 5K MIR Flickr	5000 25000
Zhang et al. 2013	JSEG algorithm	Color: 1 color feature Texture: 1 texture features Shape: 10 shape features	Decision Tree	65	-	-	Corel5K Google image	5000 5000
Bahrami and Abadeh 2014	-	-	K-nearest neighbor	30 40	33 30	31 35	Corel 5K IAPR TC-12	4999 19627
Tariq and Foroosh 2014	Divide images into 5*6 grid	Color: 18 color features Texture: 12 texture features Shape: 5 shape features	K-mean algorithm	55 45	20 19	-	IAPR-TC 12 ESP-Game	19846 21844
Zhang 2014b	the normalized cut algorithm	36-dimensional visual features for each region	Linear regression	34	24	-	Corel	5000
Zhang 2014a	the normalized cut algorithm	36-dimensional visual features for each region	Non-Linear regression (Gaussian kernel and the polynomial kernel)	25 33	41 48	-	Corel	5000
Hou and Wang 2014	-	SIFT	SVM, Spatial Pyramid and Histogram Intersection Kernels	80 84 95	-	-	Caltech-256 Corel 5k Stanford 40 actions	210 210 420
Bhargava 2014	Hessian blob detector	SURF	SVM	38	35	-	IAPR TC12	20000
Yuan-Yuan et al. 2014	-	Color: 3 color features Texture: 2 texture features	Baseline Model No-parameter Probabilistic Model	26	28	-	Corel 5K	5000
Oujaoura et al. 2014	Region growing method	Color: 1 color feature Texture: 1 texture feature Shape: 1 shape feature	SVM, Neural networks, Bayesias networks and nearest neighbor	70	-	-	ETH-80	3280
Murthy et al. 2014	-	Color : 9 color features	SVM, Discrete Multiple Bernoulli Relevance Model	36 55 56	48 25 29	-	Corel-5K ESP-Game IAPRTC-12	5000 20770 19627
Tian 2014	Normalized cut algorithm	Color: 81 color features Texture: 179 texture features Shape: 549 shape features	TSVM, Bayesian model	24	-	-	Corel 5K	5000
Majidpour 2015	-	Color: 2 color features Texture: 1 texture feature	SVM	93 64 95	-	-	image bank relate to the training set TUDarmstadt	325
Xia et al. 2015	Image's low-level features	Region area, width and high for each region	K-mean algorithm	35	44	-	IAPR TC-12	1800
SREEDHANYA and CHHAYA 2017	-	6 Features	Semi-Supervised CCA	57	46	-	LabelMe Caltech	96

The prior research demonstrates the performance that can be achieved can vary considerably, between classifiers and even with the same segmentation and feature extraction approach and dataset. It is, therefore, challenging to really understand the extent to which this approach works in practice.

Some studies have dealt with the image as one object and ignored the segmentation stage such as (Sumathi and Hemalatha 2011) (Xie et al. 2013) (Bahrami and Abadeh 2014) (Hou and Wang 2014) (Yuan-Yuan et al. 2014) (Murthy et al. 2014) (Majidpour et al. 2015) and (SREEDHANYA and CHHAYA 2017). The highest P was achieved by the studies (Sumathi and Hemalatha 2011) (Majidpour et al. 2015) and (SREEDHANYA and CHHAYA 2017) that utilized a small set of images to evaluate their performance. Indeed, it appears that as the size of the dataset increases, the retrieval accuracy decreases. This suggests results are particularly sensitive to the nature, composition and size of the dataset. This finding is also repeated in the study that employed the segmentation algorithm such as (Hidajat 2015). This is expected because an increase in the number of images that need to be analysed also leads to greater diversity in their contents, and thus the number of features needed to describe these contents will also increase. This, in turn, means that the feature extraction and comparison process to retrieve relevant images will be more complicated, and so the retrieval accuracy will be more inefficient.

On another note, (Hidajat 2015) (Sumathi and Hemalatha 2011) (Oujaoura et al. 2014) and (SREEDHANYA and CHHAYA 2017) offered good procedures for AIA and achieved high retrieval accuracy. However, these studies have been typically evaluated against datasets with a specific focus. They do not have the complexity and diversity that one might expect with a forensic investigation. The need for diversity and complexity in the forensic investigation comes from the diversity of cases that need to be solved which lead to the diversity of images contents that required to be analysed in order to find the evidence thereby solve the crime. As demonstrated above, AIA studies suffer from multiple problems. First, there is no standard annotation database for performance testing. Second, there is a disparity in system performance, because of the divergence in segmentation, features, and classifier approaches, as well as the number of images used in the assessment. Third, most studies conduct experiments using unrealistic image databases. Datasets that are unrelated to real-life complex and diverse imagery as would be expected in a forensic case. This makes it impossible to determine whether these studies would achieve a high performance in forensic image analysis.

Many commercial AIA systems that exist and have been designed by big players within the market (e.g. Google, Microsoft). However, there is little evidence or literature to suggest how well these systems work and to what extent the problems that exist within the academic literature still remain.

3 RESEARCH HYPOTHESIS AND METHODOLOGY

It is clear from the prior art that research in AIA has been undertaken independent of the forensic image analysis domain and significant progress has been made. This raised the question to what extent could existing commercial AIA systems be of benefit in forensic image analysis – where the nature of the imagery being analysed is far more complicated than has been utilized in prior studies. Therefore, the principal goal of the study was to assess the performance of these systems. An extension of this investigation was also to explore how the performance would be affected by fusion. This led to the first two experiments are:

Experiment 1: understand and evaluate the performance of commercial AIA systems using real-life imagery.

Experiment 2: determine whether a multi-algorithmic fusion approach of the aforementioned commercial systems would improve performance.

An analysis of the results from the first experiment highlighted that the annotation accompanying the datasets was not complete. This is due to missing annotations or indeed having the incorrect classified annotation in the dataset. Therefore, a further experiment was undertaken where a subset of the images was manually annotated (this included the original annotation accompanying the dataset):

Experiment 3: re-evaluate the performance based upon a more robust dataset annotation.

In order to conduct these experiments, there is a need for a dataset upon which run the experiments against. An essential requirement for the dataset was to simulate (as close as possible) image characteristics that are similar to those that would be obtained in a forensic investigation. These special characteristics include images that contain multiple objects with different sizes and orientations, irregular background, vary in quality, unconstrained illumination and different resolutions. Consequently, two publically available datasets IAPR-TC 12 (Tariq and Foroosh 2014) (Bhargava 2014) and (Xia et al. 2015) and ESP-Game (Tariq and Foroosh 2014) and (Murthy et al. 2014) were identified, because the researcher was unable to access any real-life forensic image datasets that were fully annotated. These two datasets contain various images with various characteristics, and all images in both datasets are fully annotated and thus suitable for evaluating the performance of the commercial AIA systems and the proposed approach. IAPR-TC 12 contains 19,627 images, with a resolution of 480 x 360, from locations around the world and with varied content such as places, animals, people, and birds. The ESP-Game dataset contains 20,770 images that have various images with different image sizes.

Experiment 1 Methodology

The purpose of this experiment was to determine the performance of commercial systems that able to understand the contents of the image, thereby are used as automatic image annotation systems. Several commercial providers were identified [Google Cloud Vision API, Clarifai, imagga (Imagga.com 2016), and Microsoft Cognitive Services (Computer Vision API) (Microsoft Cognitive Services 2017)]. The systems were evaluated on the two different datasets, IAPR-TC 12 and ESP-Game using a random selection of 500 images from each dataset (1000 images were used for evaluation). Images are various in their contents such as human photographs, landscapes, public places, traffic, animals, clothes, tools etc. The vocabulary size for IAPR-TC 12 and ESP-Game dataset is 153 and 755 words, respectively. Precision and Recall of per word were calculated, then Average Precision (AP), Average Recall (AR) and F-measure were used to summarize the performance.

Experiment 2 Methodology

Having established the baseline performance, it became immediately apparent that the different systems performed very differently. This led to a hypothesis of whether fusion of the systems would provide for a better degree of performance. A multi-algorithmic approach was developed that consisted of three stages: annotation extraction, normalization and fusion as illustrated in Figure 1.

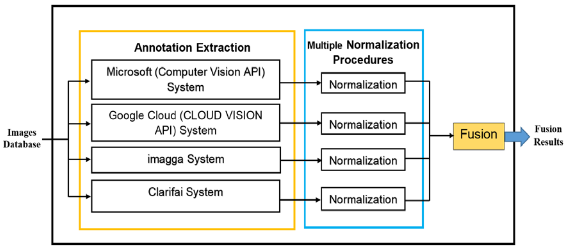
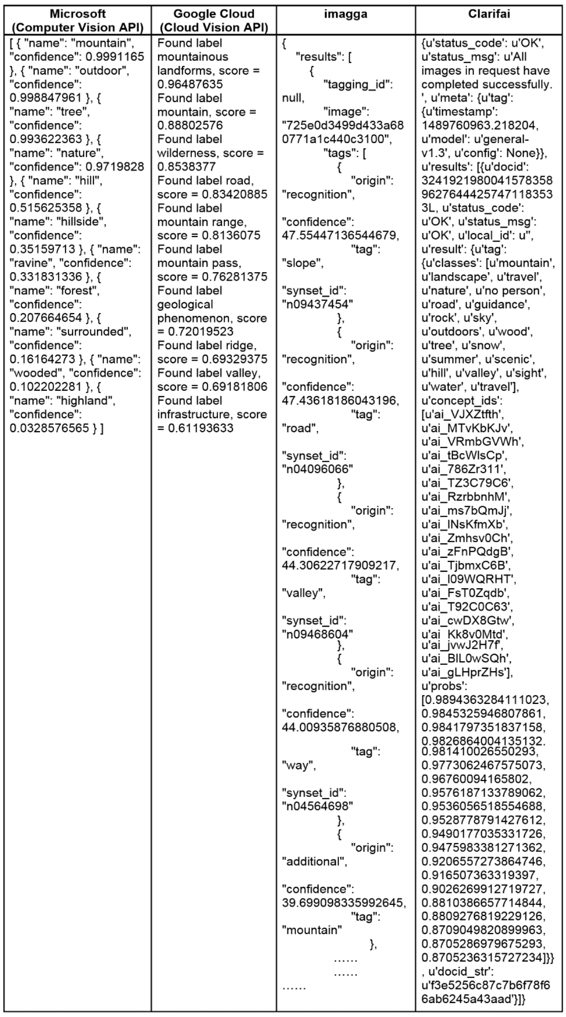


Figure 1: Block diagram of multi-algorithmic approach.

Annotation Extraction: extracts the annotations for each image in dataset through sending the image to multiple AIA systems, and then stores the result for each system. The output from each system (as illustrated in Figure 2) having a special form as compared to other annotation systems. The difference appears in the number of words that used to annotate images, the value of confidence score (probability) and in the output style of the annotations results. This leads to the problem of how to combine these different styles of annotation and express them in a unified form that can be fused to find the final annotation.

Multiple Normalization Procedures: a normalization process was required prior to fusion. The normalization process was employed to exclude all useless data and store only words and their confidence scores for each system individually in order to make confidence scores (probability) comparable to each other. The outputs were parsed and reformatted accordingly.



Microsoft (Computer Vision API)	Google Cloud (Cloud Vision API)	imagga	Clarifai
[{ "name": "mountain", "confidence": 0.9991165 }, { "name": "outdoor", "confidence": 0.998847961 }, { "name": "tree", "confidence": 0.993622363 }, { "name": "nature", "confidence": 0.9719828 }, { "name": "hill", "confidence": 0.515625358 }, { "name": "hillside", "confidence": 0.35159713 }, { "name": "ravine", "confidence": 0.331831336 }, { "name": "forest", "confidence": 0.207664654 }, { "name": "surrounded", "confidence": 0.16164273 }, { "name": "wooded", "confidence": 0.102202281 }, { "name": "highland", "confidence": 0.0328576565 }]	Found label mountainous landforms, score = 0.96487635 Found label mountain, score = 0.88802576 Found label wilderness, score = 0.8538377 Found label road, score = 0.83420885 Found label mountain range, score = 0.8136075 Found label mountain pass, score = 0.76281375 Found label geological phenomenon, score = 0.72019523 Found label ridge, score = 0.69329375 Found label valley, score = 0.69181806 Found label infrastructure, score = 0.61193633	{ "results": [{ "tagging_id": null, "image": "725e0d3499d433a680771a1c440c3100", "tags": [{ "origin": "recognition", "confidence": 47.55447136544679, "tag": "slope", "synset_id": "n09437454" }, { "origin": "recognition", "confidence": 47.43618186043196, "tag": "road", "synset_id": "n04096066" }, { "origin": "recognition", "confidence": 44.30622717909217, "tag": "valley", "synset_id": "n09468604" }, { "origin": "recognition", "confidence": 44.00935876880508, "tag": "way", "synset_id": "n04564698" }, { "origin": "additional", "confidence": 39.699098335992645, "tag": "mountain" }, …… …… ……	{u'status_code': u'OK', u'status_msg': u'All images in request have completed successfully. ', u'meta': {u'tag': {u'timestamp': 1489760963.218204, u'model': u'general-v1.3', u'config': None}}, u'results': [{u'docid': 324192198004157835896276444257471183533L, u'status_code': u'OK', u'status_msg': u'OK', u'local_id': u'', u'result': {u'tag': {u'classes': [u'mountain', u'landscape', u'travel', u'nature', u'no person', u'road', u'guidance', u'rock', u'sky', u'outdoors', u'wood', u'tree', u'snow', u'summer', u'scenic', u'hill', u'valley', u'sight', u'water', u'travel'], u'concept_ids': [u'ai_VJXZtfth', u'ai_MTvKbKJv', u'ai_VRmbGVWh', u'ai_tBcWlsCp', u'ai_786Zr311', u'ai_TZ3C79C6', u'ai_RzrbbnhM', u'ai_ms7bQmJj', u'ai_lNsKfmXb', u'ai_Zmhsv0Ch', u'ai_zFnPQdgB', u'ai_TjbmxC6B', u'ai_l09WQRHT', u'ai_FsT0Zqdb', u'ai_T92C0C63', u'ai_cwDX8Gtw', u'ai_Kk8v0Mtd', u'ai_jvwJ2H7f', u'ai_BIL0wSQh', u'ai_gLHprZHs'], u'probs': [0.9894363284111023, 0.9845325946807861, 0.9841797351837158, 0.9826864004135132. 0.981410026550293, 0.9773062467575073, 0.96760094165802, 0.9576187133789062, 0.9536056518554688, 0.9528778791427612, 0.9490177035331726, 0.9475983381271362, 0.9206557273864746, 0.916507363319397, 0.9026269912719727, 0.8810386657714844, 0.8809276819229126, 0.8709049820899963, 0.8705286979675293, 0.8705236315727234]}}, u'docid_str': u'f3e5256c87c7b6f78f66ab6245a43aad'}]}

Figure 2: Comparison between four commercial systems annotation output forms.

Fusion: the final stage of the multi-algorithmic approach was fusing the results from the four commercial systems to obtain correct and accurate annotation that describes image contents and will later be used as the query text by the investigator. The fusion stage was carried out through aggregation all annotation results that collected from four system, then the repetitions for the same word were excluded and a new probability was calculated through accumulating the probabilities that generated by the four systems for the same word as demonstrated in Table 2. After that, the final annotations were arranged in descending order depending on theirs the probabilities values in order to acquire for the final annotation of each image.

Table 2: Example of Word Repetition by Different Systems.

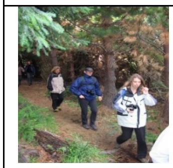
System 1	System 2	System 3	System 4	Fusion
sky	sky	sky	sky	sky
95.9426	28.5957	99.2699	96.3234	320.1316

The same datasets that utilized to evaluate the performance of the current commercial AIA systems (Experiment 1) were employed to evaluate the proposed multi-algorithmic approach performance in order to compare the performance. The results were presented in two forms. Fusion (All) based upon all annotations words and Fusion (Threshold) based upon the words having achieved a sufficient probability score of 90% or higher. This provides a focus upon the accuracy of the annotations. The Fusion (All) and Fusion (Threshold) were examined using the same two datasets that employed in the first experiment. In Fusion (All), each image was annotated with more than 50 labels. The Average Precision, Average Recall and F-measure calculated the performance.

Experiment 3 Methodology

An analysis of the results from Experiment 1 and 2 found errors within the IAPR-TC 12 and ESP-Game datasets annotations that they had been given, thereby the evaluation against with the two datasets is not fair. The two datasets were found to have incorrect and missing annotations – leading to misleading results – as many of them were incorrectly annotated. Consequently, a re-evaluation was undertaken against dataset annotation and manual re-annotation dataset for 100 images from the IAPRTC-12 dataset. In order to build manual re-annotation dataset, firstly collecting all the words that used to annotate the images based on their dataset annotation (original annotation files) in one list. After that, these images were re-annotated based on the list of words in order to create a re-annotation dataset as illustrated in Table 3. The performance of all commercial systems, Fusion (All) and Fusion (Threshold) against re-annotation and original annotation is presented.

Table 3: Examples of Image Re-annotation.

Image	Original Annotation	Re-annotation
	humans person woman landscape nature vegetation trees	Bush Face of person Grass Ground Group of persons Hat Humans Leaf Man Person Plant Tree Trees Vegetation woman

4 EXPERIMENTAL RESULTS

The following sections show the performance of the current commercial AIA systems and the proposed multi-algorithmic approach as well as the evaluation of dataset annotation.

Experiment 1

In this section, the performance of each commercial system is compared with others. These systems can obtain suitable annotation results. The findings showed that each annotation system (Microsoft, Clarifai, imagga, or Google cloud) has different levels of performance (as illustrated in Tables 4 and 5), with systems struggling more with the ESP-Game dataset. Likely, due to differing in the approaches that are used by each system to find the image annotations led to differing in the number of labels and probability values. The results also show that all systems achieved better results using IAPR-TC 12 dataset compared to the corresponding results using ESP-Game dataset. This is because the vocabulary size has many words which not match with systems annotation, and also some images in the ESP-Game dataset are small and low quality, thereby the performance of commercial systems is affected with the size of the image and its quality and this has appeared in the recent studies (Tariq and Foroosh 2014) and (Murthy et al. 2014). In addition, imagga system achieved the highest recall values for both datasets, due to a large number of words that utilized by the system to annotate each image. While, the Clarifai system achieved higher results regarding the F-measure for both datasets compared to the others systems because the number of annotations was far larger than Microsoft and Google Cloud and smaller than imagga, which made it more precise and more retrieve. Microsoft and Google cloud achieved higher precision compared with other system using IAPR-TC 12 dataset, however, their recall was low because they used a little number of words for annotation that precisely describe image content comparing with imagga and Clarifai as shown in Table 4. In addition, Microsoft's precision performance decreased in the ESP-Game dataset (as demonstrated in Table 5) because this dataset contained images with sizes less than the acceptable size that acceptable by Microsoft to find accurate label detection. Generally, the performance of these systems was low due to the quality of images that were used for evaluation, in addition to the difference between the words and its number that used by these systems and the words in dataset annotation (original annotation) that were used for evaluation.

Table 4: The Comparison of Annotation Performance for Microsoft, Google Cloud, imagga and Clarifai on IAPR-TC 12 dataset.

System Name	AP (%)	AR (%)	F (%)
Microsoft	0.38	0.31	0.34
Google cloud	0.41	0.30	0.35
imagga	0.34	0.54	0.41
Clarifai	0.36	0.52	0.43

Table 5: The Comparison of Annotation Performance for Microsoft, Google Cloud, imagga and Clarifai on ESP-Game dataset.

System Name	AP (%)	AR (%)	F (%)
Microsoft	0.23	0.18	0.20
Google cloud	0.27	0.23	0.25
imagga	0.21	0.52	0.30
Clarifai	0.29	0.45	0.35

Experiment 2

Tables 6 and 7 present the results of the multi-algorithmic approach. It was found that the performance of the multiple-algorithmic approach outperformed other commercial AIA systems against all three criteria across both datasets. Within a forensic image analysis context, the average recall (AR) is more important than average precision (AP), as it is preferably for an artefact to be identified than missed, even if this results in an investigator having to examine more images. Fusion (All) based recall rates of 76-78% against a single-classifier with the best result of 54% shows a significant improvement. Regarding the average precision (AP), the highest value achieved by Google cloud was 41% that annotates image approximately 15 words, however, Fusion (All) achieved 35% despite it annotated the image with more than 50 tags as an average. Furthermore, Fusion (Threshold) that annotates the image with more than 20 tags achieved high average precision (AP) for both datasets than the other AIA systems. Moreover, the precision of the Fusion (Threshold) is greater than the precision of Fusion (All) results, because there is an inversely proportional between the number of words and accuracy.

Table 6: The Comparison of Annotation Performance for Fusion (All) and Fusion (Threshold) on IAPR-TC 12 dataset. (Red color refers to the superiority of the proposed approach).

System Name	AP (%)	AR (%)	F (%)
Fusion (All)	0.35	**0.76**	**0.48**
Fusion (Threshold)	**0.43**	**0.58**	**0.49**

Table 7: The Comparison of Annotation Performance for Fusion (All) and Fusion (Threshold) on ESP-Game dataset. (Red color refers to the superiority of the proposed approach).

System Name	AP (%)	AR (%)	F (%)
Fusion (All)	**0.32**	**0.78**	**0.46**
Fusion (Threshold)	**0.37**	**0.50**	**0.42**

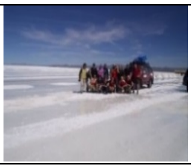
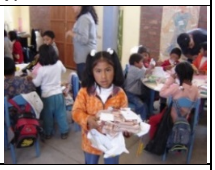
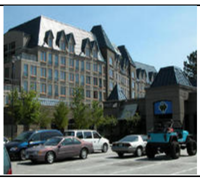
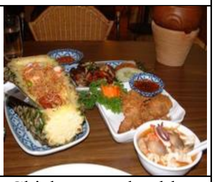
Validating the semantic retrieval performance of the multi-algorithmic fusion approach, Precision, Recall and F-measure were employed to evaluate the single word retrieval performance. The retrieval performance was tested separately based on dataset annotation, Fusion (Threshold) and Fusion (All), and the F-measure (F) values were 72.4%, 84.0% and 77.5%, respectively as shown in Table 8. These results showed the superiority of the multi-algorithmic fusion approach over original annotation (IAPR-TC 12 dataset); despite some of the images were very small, low in contrast or have part of the requested object. In addition, the image object itself differs in shape, color, size, location and direction in each image. The Fusion (All) annotation achieved the lower average precision, because it retrieves some images that have objects related with the tested word; however, it successfully retrieved all images that have the tested words in their content, and its Average Recall (AR) is 98%. This means that the proposed approach will help the investigator to retrieve all requested evidence from the images dataset; thereby it will facilitate the process of identifying and solving the crimes.

Table 8: The Retrieval Performance Based on One Word Queries. (Red color refers to the superiority of the proposed approach).

	Dataset annotation		Fusion (Threshold)		Fusion (All)	
Words	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
car	97.7	86	96	96	75.3	100
food	100	69	91.4	76.1	78.8	97.6
dog	100	100	92.3	92.3	75	92.3
Flower/ rose	100	1.25	85.7	60	75	100
cold	100	27.7	83.3	55.5	51.5	94.4
bicycle	100	33.3	100	100	66.6	100
bed	100	85.7	77.7	100	63.6	100
boy	100	51.6	65.7	74.1	27.6	100
Average	99.7	56.8	86.5	**81.7**	64.1	**98**
F	72.4		**84.0**		**77.5**	

In addition, the comparison between dataset annotation and Fusion (Threshold) annotations results indicates that the original annotation lost some words and does not provide synonyms or substitute words that describe the same image content. The proposed approach predicted annotations (words) for the images better than dataset annotation (original annotation) in three issues. Firstly, it is more accurate in describing image content. Secondly, the number of words that describe the image by the proposed approach is greater than dataset annotation, as well as the multi-algorithmic describes all image contents efficiently that will help on not miss any object in the image. Thus, the proposed approach can solve the problem of poor annotation (images are not annotated with all relevant keywords) and overcome the limitations above in AIA studies. Finally, it offers many synonyms and describes the whole image content as illustrated in Table 9.

Table 9: Examples of Fusion Annotation Matching with Ground Truth Annotation for Two Datasets (APR-TC 12 and ESP-Game).

APR-TC 12 Dataset		
Image		
Original Annotation	Humans, group of persons, landscape nature, sky	Humans, person, child, child girl, man made, floor
Fusion Annotation	Snow, sky, winter, ice, cold, outdoor, landscape, travel, outdoors, water, beach, people, leisure, vacation, frosty, vehicle, froze, recreation, frost, weather	People, group, education, class, child, person, adult, classroom, boy, school, man, room, teacher, woman, indoor, wear
ESP- Game Dataset		
Image		
Original Annotation	Car, building	Chicken, meal, table, bowl, food, white, Asian, dinner
Fusion Annotation	Building, sky, road, street, town, downtown, architecture, city, travel, outdoor, urban, house, tourism, old, outdoors, car, modern, horizontal, facade	Food, meal, plate, dish, table, cuisine, lunch, restaurant, dinner, meat, delicious, sauce, vegetable, healthy, tasty, cooking, hot, indoor, epicure, refreshment, no person

Experiment 3

Correcting for errors or missing in the annotation that came with the dataset shows the overall precision has improved across the board (as illustrated in Figure 3), with Fusion (Threshold) achieving the highest performance. This means that the re-annotation dataset enables significantly more precise and true results than dataset annotation (IAPRTC-12 dataset) because the re-annotation dataset addressed the missing and wrong annotations issues.

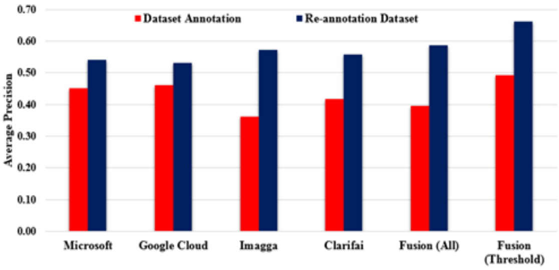


Figure 3: Average precision of the six systems with two different annotation datasets.

For Average Recall values, opposite results were obtained (as presented in Figure 4), because the re-annotation dataset is more precise (inverse relationship between precision and recall). However, the AR of the Fusion (All) in the re-annotation dataset is still higher than the other online existing AIA systems because Fusion (All) includes all annotations that collected from all systems. Generally, the F-measure value of Fusion (All) is higher than the other AIA systems and Fusion (Threshold) especially using re-annotation dataset as shown in Figure 5. The issue re-annotating introduces is the expansion in the number of annotations listed for each image. Consequently, the Fusion (Threshold) is negatively impacted. The results of this investigation show that the Fusion (All) and Fusion (Threshold) in all metrics were higher than other systems regardless the dataset validity used for evaluation – supporting the use of a multi-algorithmic approach.

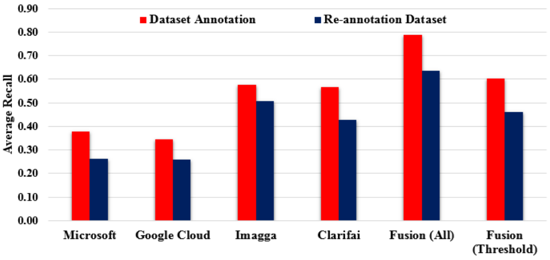


Figure 4: Average recall of the six systems with two different annotation datasets.

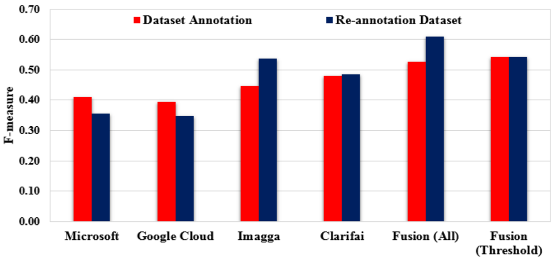


Figure 5: F-measure of the six systems with two different annotation datasets.

5 DISCUSSION

The evaluation of different commercial AIA systems (as illustrated in experiment 1) revealed the performance of these systems contrast against the same or different datasets. This is because these systems describe a given image in three different ways: 1) only on the main objects in the image; 2) the same object with different words (synonyms), and 3) the main objects, synonym and the general description of the whole image content. In addition, the results showed the highest performance for all systems was achieved by using IAPR-TC 12 dataset compared to the corresponding results using ESP-Game dataset, it is expected because the ESP-Game dataset contains some small and low-quality images, in addition to the small number of vocabulary that used for annotations. This means that the performance of the systems is affected negatively by the quality and size of the image. The second conducted experiment results showed the performance of AIA is improved through the fusion of many systems. Image annotation results from an individual commercial AIA system constructively improved through the combining between results of multiple AIA systems. This because of the increase in the number of annotations, collects alternatives words for the same object (synonym), describe whole image content as well as its objects, in addition to increasing the reliability of the words that have high probability score because they are repeated by different systems. The proposed approach is able to retrieve all images that have the text query (tested word) in their content successfully and average recall rate was 98%, as well as improved image annotation and solved the problem of poor annotation (images are not annotated with all relevant keywords). The last conducted experiment results highlighted that usage re-annotation dataset improved all systems precision performance because finding some mistakes in dataset annotation. Additionally, the proposed approach achieved better performance than the rest of the systems regardless of the dataset that used for evaluation.

However, the use of publically available annotation systems introduces some operational limitations. Firstly, some of these systems such as Microsoft Vision API take a copy of the image in order to improve its system performance. Secondly, there is a variety of forensic images evidence that has been captured by different devices; some of them are often poor quality and highly variable in size and content. Thus, the precision of annotation that obtained from available commercial annotation systems affected by several factors such as image clarity, image size, and size and direction of an object in the image. Consequently, there is a need to explore and evaluate a range of pre-processing procedures to introduce the necessary privacy required and tackle image factors.

6 CONCLUSIONS

In this paper, the performance of existing commercial AIA systems, as well as the proposed multi-algorithmic approach were evaluated. The experimental results using two datasets show that the proposed method outperforms the existing AIA systems. The proposed method annotated the image with many correct and accurate words that reflecting image content and will later improve the retrieval performance. The results also argued that the proposed approach improved the efficiency and accuracy of the image annotation comparable to the state of the art works.

Future work, however, needs to seek, explore and evaluate a range of pre-processing procedures to achieve the necessary privacy. Furthermore, additional research in image enhancement should be conducted to improve image quality that would improve the annotation systems performance, thereby improving the performance of the multi-algorithmic approach.

REFERENCES

Bahrami, S. & Abadeh, M.S., 2014. Automatic Image Annotation Using an Evolutionary Algorithm (IAGA). *2014 7th International Symposium on Telecommunications (IST'2014)*, pp.320–325.

Bhargava, A., 2014. An Object Based Image Retrieval Framework Based on Automatic Image Annotation.

Calrifai, 2018. API | Clarifai. Available at:

https://www.clarifai.com/api.

Al Fahdi, M. et al., 2016. A suspect-oriented intelligent and automated computer forensic analysis. *Digital Investigation*, 18, pp.65–76. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1742287616300792.

Forensicsciencesimplified.org, 2016. Forensic Audio and Video Analysis: How It's Done. Available at: http://www.forensicsciencesimplified.org/av/how.html.

Google Cloud Platform, 2017. Vision API - Image Content Analysis | Google Cloud Platform. Available at: https://cloud.google.com/vision/ [Accessed April 10, 2017].

Hidajat, M., 2015. Annotation Based Image Retrieval using GMM and Spatial Related Object Approaches. , 8(8), pp.399–408.

Hou, A. & Wang, C., 2014. Automatic Semantic Annotation for Image Retrieval Based on Multiple Kernel Learning. , (Lemcs).

Imagga.com, 2016. imagga - powerful image recognition APIs for automated categorization &amp; tagging. Available at: https://imagga.com/.

Inoue, M., 2004. On the need for annotation-based image retrieval. *in: IRiX'04: Proceedings of the ACM-SIGIR Workshop on Information Retrieval in Context, Sheffield, UK*, pp.44–46. Available at: https://pdfs.semanticscholar.org/af7c/e5f0531a660782535dd954445c530a0c34b0.pdf [Accessed June 23, 2017].

Lee, J. et al., 2011. Image Retrieval in Forensics: Application to Tattoo Image Database. *IEEE Multimedia*.

Li, Z. et al., 2012. Combining Generative/Discriminative Learning for Automatic Image Annotation and Retrieval. *International Journal of Intelligence Science*, 2(3), pp.55–62. Available at: http://www.scirp.org/journal/PaperDownload.aspx?DOI=10.4236/ijis.2012.23008.

Majidpour, J. et al., 2015. Interactive tool to improve the automatic image annotation using MPEG-7 and multi-class SVM. In *2015 7th Conference on Information and Knowledge Technology (IKT)*. IEEE, pp. 1–7. Available at: http://ieeexplore.ieee.org/document/7288777/.

Microsoft Cognitive Services, 2017. Microsoft Cognitive Services - Computer Vision API. Available at: https://www.microsoft.com/cognitive-services/en-us/computer-vision-api [Accessed April 10, 2017].

Murthy, V.N., Can, E.F. & Manmatha, R., 2014. A Hybrid Model for Automatic Image Annotation. In *Proceedings of International Conference on Multimedia Retrieval - ICMR '14*. New York, New York, USA: ACM Press, pp. 369–376. Available at: http://dl.acm.org/citation.cfm?doid=2578726.2578774.

Oujaoura, M., Minaoui, B. & Fakir, M., 2014. Combined descriptors and classifiers for automatic image annotation.

Redi, J.A., Taktak, W. & Dugelay, J.-L., 2011. Digital image forensics: a booklet for beginners. *Multimedia Tools and Applications*, 51(1), pp.133–162. Available at: http://link.springer.com/10.1007/s11042-010-0620-1.

Singh, A., 2015. Exploring Forensic Video And Image Analysis. Available at: https://www.linkedin.com/pulse/exploring-forensic-video-image-analysis-ashish-singh.

SREEDHANYA, S. & CHHAYA, S.P., 2017. Automatic Image Annotation Using Modified Multi-label Dictionary Learning. *International Journal of Engineering and Techniques*, 3(5). Available at: http://www.ijetjournal.org [Accessed March 9, 2018].

Sumathi, T. & Hemalatha, M., 2011. A combined hierarchical model for automatic image annotation and retrieval. In *2011 Third International Conference on Advanced Computing*. IEEE, pp. 135–139. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5329188.

Tariq, A. & Foroosh, H., 2014. SCENE-BASED AUTOMATIC IMAGE ANNOTATION. , pp.3047–3051.

Tian, D., 2014. Semi-supervised Learning for Automatic Image Annotation Based on Bayesian Framework. , 7(6), pp.213–222.

Tian, D., 2015. Support Vector Machine for Automatic Image Annotation. , 8(11), pp.435–446.

Worthington, P., 2015. One Trillion Photos in 2015 - True Stories. Available at: http://mylio.com/true-stories/tech-today/one-trillion-photos-in-2015-2 [Accessed September 26, 2017].

Xia, Y., Wu, Y. & Feng, J., 2015. Cross-Media Retrieval using Probabilistic Model of Automatic Image Annotation. , 8(4), pp.145–154.

Xie, L. et al., 2013. A Two-Phase Generation Model for Automatic Image Annotation. In *2013 IEEE International Symposium on Multimedia*. IEEE, pp. 155–162.

Yuan-Yuan, C. et al., 2014. A hybrid hierarchical framework for automatic image annotation. In *2014 International Conference on Machine Learning and*

Cybernetics. IEEE, pp. 30–36. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7009087.

Zhang, D., Monirul Islam, M. & Lu, G., 2013. Structural image retrieval using automatic image annotation and region based inverted file. *Journal of Visual Communication and Image Representation*, 24(7), pp.1087–1098. Available at: http://dx.doi.org/10.1016/j.jvcir.2013.07.004.

Zhang, N., 2014a. A Novel Method of Automatic Image Annotation. *Computer Science & Education (ICCSE), 2014 9th* ..., (Iccse), pp.1089–1093. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6926631.

Zhang, N., 2014b. Linear regression for Automatic Image Annotation. *Computer Science & Education (ICCSE), 2014 9th* ..., (Iccse), pp.682–686. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6926548.