J Ambient Intell Human Comput (2014) 5:909–918
DOI 10.1007/s12652-014-0248-5

ORIGINAL RESEARCH

On user authentication by means of video events recognition

Luigi Catuogno · Clemente Galdi

Received: 6 June 2014 / Accepted: 8 October 2014 / Published online: 15 October 2014

Abstract Graphical password schemes have been widely analyzed in the last couple of decades. Typically such schemes are not resilient to adversaries who are able to collect a considerable amount of session transcripts, and can process them automatically in order to extract the secret. In this paper we discuss a possible enhancement to graphical passwords aiming at making infeasible to the attacker to automatically process the collected transcripts. In particular, we investigate the possibility of replacing static graphical challenges with on-the-fly edited videos. In our approach, the system challenges the user by showing her a short film containing a number of pre-defined pass-events and the user replies with the proof that she recognized such events. We present a proof-of-concept prototype, FilmPW, and discuss some issues related to event life-cycle management. Our preliminary experiments show that such an authentication mechanism is well accepted by users and achieves low error rates.

Keywords Graphical password · Authentication · Human cryptography

L. Catuogno (✉)
Dipartimento di Informatica, Universitá degli Studi di Salerno, Via Giovanni Paolo II, 132, I-84084 Fisciano, SA, Italy
e-mail: luicat@dia.unisa.it

C. Galdi
Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione, Universitá degli Studi di Napoli "Federico II", Compl. Univ. Monte S. Angelo, Via Cinthia, I-80126 Napoli, Italy
e-mail: clemente.galdi@unina.it

1 Introduction

In recognition based graphical password schemes, the system challenges the user by showing her an arrangement (scene) of small icons (usually chosen from a pre-defined finite "portfolio") and successfully authenticates her if she provides a response that proves she correctly distinguished the icons which belong to her secret (pass-icons) amongst the others. We argue that such schemes are particularly prone to *shoulder surfing* attacks that leverage malicious agents installed on the user terminal to gather any useful information regarding the authentication process such as the session transcript or statistical data about the user's behavior.

There are several ways to do this: Skimmers and quick cameras can be placed upon the console of an ATM as well as key loggers and spy-ware of any kind can be installed on the computer the user employs to access the authenticator system, in order to collect the data transferred to/from security tokens, to record which keys were pressed by the user, what appeared on the screen during the authentication session and so on.

Such type of attacks targeting the user device can also defeat authentication schemes that are based on complex underlying security mechanisms. Consider, for example, the two factor variant of PassFaces(Real User Coorp. 1998) in which a mutual endpoint authentication, by means of the SSL protocol, takes place along with the graphical password scheme. In this case, the presence of the SSL protocol guarantees the integrity and authenticity of the messages. On the other hand this measure does not prevent a spy-ware from gathering data exchanged through the graphical authentication channel.

Once the adversary has collected enough information, she tries to extract the user's secret by automatically

processing such data. To perform such extraction the adversary should have a way to model the sessions so that processing of data results computationally affordable.

For example, (Golle and Wagner 2007) presented an attack to a recognition-based scheme proposed in (Weinshall 2006), in which, using data coming from a certain number of previously recorded authentication sessions (transcripts), the adversary creates a system of boolean equations whose solutions (by means of a SAT solver) lead to the user's secret discovery. In such an attack the adversary is supposed to have the ability of identifying each single pass-icon and label it with a different variable name, so that the matrix of objects shown in the visual challenge is mapped onto a set of equations.

We point out that this is not an unrealistic possibility, though it depends on scheme's setup. On one hand, if challenges are formed as HTML/XML documents, and sent to the prover over the network, the adversary could recognize each pass-icon by related meta-data such as its file name or content.

On the other hand, in several settings (*e.g.*, in the ATM scenario), the adversary could be just enabled to take a screenshot of the challenge "as it appears", without having any access to meta-data and other information. In this case, the adversary could still extract information about the challenge composition (which pass-icons are present and in which position), by leveraging object detection techniques (Jones and Viola 2001). Following this approach, the adversary is able to recognize single objects within the screenshot and to group them in such a way that objects which fall in the same group are, with a high probability the same pass-icon and can be mapped to the same variable name. Such tools are capable to make a classification choice very fastly (few seconds), though an adequate preliminary training phase (that can be also done off-line) is required. Furthermore, this approach could be still suitable in the case in which the system does not simply pick the pass-icons from a fixed image portfolio but, instead, it composes the challenge using dynamically distorted versions of the original pass-icons or arbitrary images similar to the original pass-icons, according certain criteria: typically the subject (*e.g.*, any ball instead of *that* ball, and so on).

Once the equations system is solved, the adversary obtains a binary result-vector X where the value of its i-th component indicates whether or not the i-th object belongs to the user secret. A similar attack is discussed in (Catuogno and Galdi 2008, 2010, 2014).

Two main approaches can be adopted in order to prevent the usability of this type of automatic attacks. The first one consists in limiting the information released by the scheme at each session. If the released information is carefully selected, the computational effort required by the adversary in order to complete the attack might make the attack unfeasible.

However, this approach, might have negative effects on usability.

The second approach consists in building schemes in which user and system interact in a way such that an automatic agent is unable to efficiently capture the communication content (*e.g.,* by tracking eye movement or through tactual stimulations). Two factor schemes also fall in this category. Although such schemes are quite effective, they leverage on special devices (often costly or not well-accepted) that cannot be deployed in every usage scenario.

To summarize, we can identify the following problems: (1) the "atomic" components of the majority of graphical password schemes, scenes, pass-icons and user actions, are often (too) easily distinguishable and collectable for an automatic procedure; (2) due to the fact that such authentication schemes are designed to be computed by humans, they result to be breakable with high probability by an adversary which has collected and efficiently modelled a sufficient number of transcripts.

Our Contribution One of the issues that arise from the above discussion is that computers can "easily" recognize (static) images. Such property allows the construction of automatic shoulder surfing attacks based on the analysis of sequences of authentication sessions. It is possible to generalize such information extraction, to some extent, to the case of videos in the sense that, given a static image, it is possible to check whether or not a video contains a scene that "resembles" somehow the one shown by the static image.

In this paper we analyze for the first time the possibility of fully exploiting the potential of the human mind for recognizing specific *concepts or actions* independently from the actors or the context in which such actions or concepts are represented. The key observation is that a person can easily recognize actions or concepts within a video. This is true even in the case in which the specific video has not been shown to the user before. On the other hand there exists, to the best of our knowledge, no software that is able to automatically classify arbitrary videos based on the (arbitrary) actions or concepts shown therein.

All previously known graphical password schemes grant authentication if the user is able to recognize or reproduce *specific pictures* among others, that corresponds to recognize some *shapes* among others. The novelty of our approach consists in binding the success of the authentication to the *"meaning" of a object* instead of its *shape*.

We have used such an observation for designing an authentication mechanism in which the user recognizes actions or concepts within a sequence of short videos. We have implemented a prototype that allows the execution of authentication sessions using our new video-based scheme. Although the prototype is simple, we have conducted some preliminary usability tests. In our experience the new

authentication mechanism has good acceptance rates among the users.

2 Related works

In this section we briefly review the related works in the field of graphical passwords and video event recognition.

2.1 Graphical passwords

Graphical passwords (Suo et al. 2005) are a possible security enhancing and user-friendly alternative to the old-fashioned password authentication scheme, relying on the fact that humans remember images better than characters (Grady et al. 1998; De Angeli et al. 2005). Amongst several attempts to enhance security and usability of old-fashioned text-based password schemes (Haller 1994; McDonald et al. 1995; Blundo et al. 2004; Ciaramella et al. 2006), new "human affordable" authentication protocols (Matsumoto 1996; Blonder 1996; Hopper and Blum 2001), including graphical passwords, have been studied since the early 90s. However, graphical passwords have known a major development in the last decade, leveraging the increasing availability of human-computer interaction technologies.

Graphical Password schemes are frequently classified according the cognitive process they rely on and are mainly divided in three classes (Wiedenbeck et al. 2005): *Pure Recall-based*, *Cued Recall-based* and *Recognition-based* schemes.

In *Pure Recall-based* schemes (Jermyn et al. 1999), the challenger requests the user to paint, on an empty canvas/grid, a pre-defined two dimensional picture in the same way she did during the registration phase. *Cued Recall-based* schemes, also known as *Click-based schemes* (Blonder 1996; Wiedenbeck et al. 2005), challenge the user showing her an image (scene) and ask the user to click on some previously chosen regions, according to a certain sequence. Scenes, regions and sequences of clicks are selected at user registration time as the secret. In (Maetz et al. 2009), an approach somehow related to ours is presented. The user creates her secret by "telling a story" via choosing a background scene and placing a certain set of pass-icons on it. In order to be authenticated, the user has to reproduce a certain sequence picked from the whole story. *Recognition-based* schemes (Dhamija and Perring 2000; Sobrado and Birget 2002; Jensen et al. 2003; Real User Coorp. 1998) rely on the users ability to recognize a set of' "pass-icons" or "pass-images" included, along with other "decoys", in a randomly arranged matrix shown as challenge by the system.

In Recall-based schemes, an actively studied security threat is the *predictability* of the secret. Some works address the fact that some click-points (hot-spots) can be chosen by the user, as parts of her secret, more likely than others (Thorpe and van Oorschot 2007; Salehi-Abari et al. 2008; Bicakci et al. 2009), giving to an adversary some advantages in *guessing* the secret. On the other hand, instead of trying to guess the user secret, an adversary could try to capture it by observing the legitimate user "over her shoulder" when she logs-in to the system, in order to learn the secret. Notice that, the objective of early schemes was ergonomics, and did not feature any effective countermeasures against a malicious observer, who, in some case, could learn the user secret just observing a single or few authentication sessions.

To overcome these threats, the following two strategies could be adopted. The first strategy aims, as in classical text-based password schemes, to make difficult to the adversary to realize and understand what happens during the user log-in. In (Harada et al. 2006; Hayashi et al. 2008), the system shows the user a distorted version of her pass-icons, leveraging on the assumption that being aware of the original pictures, the legitimate user is able to recognize them, whereas the adversary cannot.

In (Gao and Liu 2009), pass-icons are displayed as CAPTCHAs, in order to make infeasible their automatic recognition. However, recently, several threats to CAPTCHAs have been documented (Yan and El Ahmad 2008; Golle 2008; Li et al. 2010; Bursztein et al. 2011).

In (Kumar et al. 2007; De Luca et al. 2009; Lanat et al. 2013), the system gets the user response by means of a device that tracks her eye movements, making impossible to the adversary to capture user actions. The Undercover (Sasamoto et al. 2008) system, integrates the graphical challenge with undetectable tactile stimulations by means of a special human-computer interface device.

The second strategy assumes that the adversary is able to extract the transcripts of a certain number of authentication sessions and aims to lower the advantage that she can obtain in guessing the secret by processing the collected information. Schemes described in (Roth et al. 2004; Wiedenbeck et al. 2006; Weinshall 2006; Jameel et al. 2006; Catuogno and Galdi 2008) replace the simple image recognition with a sort of *cognitive game* the user has to play in order to be authenticated and that she can easily answer correctly to the challenge if she is aware of the secret. Although these schemes effectively contrast attacks mounted by transitory human observers (few memory, poor computation capabilities), they are still vulnerable to attacks moved by camera-equipped adversaries, i.e., an adversary that can record an unbounded number of interactions of the user with the terminal. In such cases, the adversary is able to extract the secret, with high probability, by trivially processing a dozen of transcripts (Catuogno and Galdi 2008). In particular, the

authors in (Golle and Wagner 2007) show how using a SAT solver (Tompkins and Hoos 2004) can break the scheme in (Weinshall 2006) in few minutes, having collected just six-seven transcripts. Extracting the user secret by processing an arbitrary number of sessions in the HB protocol family (Hopper and Blum 2001) has proven to be hard, though, to date, we do not know any graphical password scheme that implements one of these protocols. User authentication through music recognition is investigated in (Gibson et al. 2009).

2.2 Video event recognition

The field of Video Event Recognition is an extremely active research area. The main motivations for research in this field are related to video surveillance (e.g., field monitoring, interaction among persons in public areas etc.), and biometric authentication, e.g., face recognition.

In particular we are interested in the process of semantic video annotation (Ryoo et al. 2010; Snoek and Worring 2005; Brezeale and Cook 2008; Lavee et al. 2009; Ko 2008; Hoque et al. 2013), that is the possibility of automatically classifying videos according to the actions performed therein. Key solutions in this field consist in considering the classification problem as a (either supervised or unsupervised) learning problem in which the system is trained to identify specific sets of predetermined actions. An example is the DARPA Mind's Eyes program whose goal is the detection and annotation of the actions occurring in unstructured videos and expressed by 48 English verbs such as, *walk, carry, dig*, etc.

A slightly different approach is the one presented by (Merler et al. 2012) in which real-world (YouTube-like) videos are classified by a system starting from thousands of labelled images, describing a total of 280 concepts. Clearly there exits a trade-off between generality and precision of classification as the higher is the former, the lower is the latter. To the best of our knowledge, however, currently there exists no system that allows the automatic classification of *arbitrary* videos.

3 Preliminaries

The key observation we use in this paper is the following: it is common knowledge that each human can easily recognize a given action or concept in every (reasonable) video showing some actor performing the action or representing the concept. The only prerequisite for this recognition to be feasible is that the person needs to *know* what she is looking for/at. In order to understand the latter statement we need to make a distinction between actions and concepts.

In general, we might better specify an action by specifying (a) the subject that performs the action; (b) the verb that identifies the action and (c) the object on which the action is executed. Notice, however, that an action is only described by the verb while providing the subject and the object restricts the set of possible scenarios described by a given phrase.

For example, we might state that the action *eat* describes all the actions in which *somebody eats something*. We can restrict the set of actions by specifying either the subject, e.g., *man eats something*, or the object, *somebody eats donut*, or both, *man eats donuts*. If we assume that a person knows the meaning of the verb *eat* and the meaning of the word *donut*, such a person can easily recognize the concept *eat a donut* in different contexts, e.g., *man eats a chocolate donut in a diner* or *woman eats a strawberry donut while pirouetting on the street*, and so forth.

A different notion of knowledge might be required for *concept recognition*. We classify concepts as *universally recognizable* or *locally recognizable*. An example of the former type of concepts is *happiness*. Indeed, if the feeling is evident from the video, a person is recognized as happy independently from the culture of the observer. In general, however, a concept might be related to some knowledge that is typical of the culture of the person observing the video. A simple example might be the concept *wedding*. The wedding ceremony (if any) differs from place to place. Indeed, every culture has its own schemas for such events, e.g., dressing code, ceremony, location, reception, etc. For this reason, the specification of a locally recognizable concept might need some type of restrictions in order to be properly recognizable.

4 FilmPW at glance

In this section we present the intuition of a novel graphical authentication scheme: FilmPW, which aims to overcome the drawbacks discussed above, the pass-icons based recognition schemes suffer from.

Unlike the majority of recognition based schemes, FilmPW challenges the user with a seamless flow of information (a video stream) instead of visualizing static scenes composed of a finite set of pass-icons. In such a way, FilmPW aims to make difficult the automatic recognition of the symbols composing the challenge. Roughly speaking, FilmPW asks the user to recognize a sequence of secret actions or concepts, the *pass-events*, that are included in a short film, by typing a set of alphanumeric strings corresponding to the recognized pass-events.

At registration time, the user choses, as secret, a set of scenes which can recall personal memories, dreams or general concepts. Suppose, for example, the set of the

user's pass-event includes a scene in which a cat walks on the roof, a man eating an ice cream and a wedding ceremony. Moreover, for each pass-event, the user provides an associated set of response symbols, (*e.g.,* an alphanumeric string, a number, a sequence of the pointing device) used to describe each event in the response; in our example, the user maps the mentioned events, respectively, to the responses `'11'`, `'Mich'` and `'mayday'`. During the authentication process, the system challenges the user visualizing a short video edited using a certain number of events chosen from the user's pass-events plus some distracting events. In our example, the system shows the cat walking on the roof a man eating an ice cream interposed with some irrelevant scenes. In order to be authenticated, the user has to type the correct response: `'11Mich'`.

5 The authentication scheme

In this section we describe the authentication scheme. We first identify the different components we will deal with. We then identify the different actors that are part of the architecture of our authentication system. Finally we describe the phases composing our protocol.

5.1 Components

The building blocks of our authentication protocol are the following.

Events. In FilmPW, events are short video fragments in which something happens. Practically, any event is a file stored in a database which can be identified by means of a set of tags. We denote the events database with $\mathcal{S}$ and with $\mathcal{T}$ the set of all possible tags. Each event $e_i \in \mathcal{S}$ is indexed by a set of tags $T(e_i) = \{t_1, \ldots, t_n\}$ (with $T(e_i) \subset \mathcal{T}$), that give information about its content or meaning. Tags are assigned to events at creation time, *e.g.* when the clip is edited and added to the database. In our experiments, $\mathcal{T}$ is composed of words of the natural language. In order to chose a pass-event, the user provides such a set of tags. Each event $e_i \in \mathcal{S}$ can be retrieved by running a query $q(t_1, \ldots, t_k)$ (with $t_1, \ldots, t_k \in \mathcal{T}$). More precisely, the query q returns a set of events $\{e_j | t_1, \ldots, t_k \subseteq T(e_j)\}$. In our example, tags assigned to the pass-event "a cat walks on the roof" include (but are not limited to) "cat", "walk" and "roof". We point out that it is possible that the same query $(t_1, \ldots, t_k)$ can identify different events as well as the same event $e_i \in \mathcal{S}$ could be identified by different queries.

Responses. During the registration phase, the user selects a set of symbols that should be used in composing the responses to the challenges. We denote with $\mathcal{R}$ the space of response symbols. In our prototype, $\mathcal{R}$ is a set of arbitrary alphanumeric strings. However, in order to keep the response space sufficiently large, it would be preferable to make different choices for $\mathcal{R}$ such as a set of regular expressions. Moreover, the system could offer different alternatives for $\mathcal{R}$ according to the device that will be used to perform authentication.

Secrets. The secret $S(u)$ of user u is denoted with the triple (Q, R, M) that is composed of: the set of queries $Q = \{q_i(t_{i_1}, \ldots, t_{i_k})\}$, the set of response symbols $R \subseteq \mathcal{R}$ and a mapping $M : Q \rightarrow R$. Additionally, we define the set of user's pass-events $E(u) = \{e_i \in \mathcal{S} | e_i \in q_j | q_j \in Q\}$ and, the extended mapping $M' : E(u) \rightarrow R$ which $\forall q_i \in Q$ maps any event $e_j \in E(u)$ to the response symbol $r_i \in R$ if $e_j \in q_i$. In other words, M' maps each event $e_j \in \mathcal{S}$ to the same response symbol r_i if the tags of its associated query q_i form a subset of $T(e_j)$.

Challenges. A challenge $C(u)$ is a 30 seconds length video the system edits in order to authenticate the user u whose secret is $S(u) = (Q, R, M)$. $C(u)$ is a choice of n events $\{c_{i_1}, \ldots, c_{i_n}\}$ picked from $\mathcal{S}$ such that $k = \lfloor n/2 \rfloor$ of them belong to $E(u)$ (the pass-events $c_{i_1}, \ldots, c_{i_k}$) while the rest do not (the distracting events $c_{i_{k+1}}, \ldots, c_{i_n}$). Each pass-event $c_{i_j}, (1 \leq i_j \leq k)$ is randomly chosen from the result of a different query in Q. Whereas the remaining $n - k$ distracting events are randomly chosen from $\mathcal{S}$ and selected for the video editing process if they do not match any query in Q.

5.2 The architecture

The authentication protocol is executed by the following players.

The User. In order to access the system, the user needs to pass through a registration phase in which the system gives her a unique user id u and helps her in creating a secret $S(u) = (Q, R, M)$. The communication between the user and the registration service is assumed secure, and the registration service itself is trusted. However, the communication channel between user and authenticator is untrusted.

The Authenticator. The authenticator stores the users' secrets and interacts with the user according to the authentication protocol described below. It receives the request for authentication by user u and recognizes her user id, issues a challenge to the user and receives and validates her response. For the sake of challenge editing, the authenticator makes the required queries to the events database manager. At the end of the authentication session, the authenticator responds with either **1** or **0** depending on whether the authentications succeeds or not. The authenticator is trusted and the communication channel with the database manager (at this stage of our work) is also trusted.

The Event Database Manager(EDBM). The EDBM is a DBMS that stores the events in $\mathcal{S}$ indexed through tuples in $\mathcal{T}$. The EDBM receives the queries by the authenticator and returns a single event, randomly chosen among the ones that match the query. As required by the authentication protocol, the EDBM also features untagged queries, that return a set of randomly chosen events from $\mathcal{S}$. At the current stage of our work, the EDBM is assumed trusted and the tagged database $\mathcal{S}$ is provided at the setup of the system "as is". Nevertheless, as we discuss later, this component raises some non trivial issues (see Sect. 6) that will be addressed in a future work.

5.3 The protocol

The authentication protocol consists of two phases. The first one, the Registration phase, is executed only once for each user.

Registration. During the registration phase the user creates her secret $S(u) = (Q, R, M)$. To this end, the user first selects her pass-events by providing, for each event, a set of keywords that are used as tags to form the queries in Q. In this way, the user implicitly selects as pass-events all events $e_i \in \mathcal{S}$ that match the formed query. Secondly, the user implicitly builds the set R and the mapping M by choosing, for each query in Q, the corresponding response symbol in $\mathcal{R}$. Note that the user can also map the same event to different response symbols.

At the end of the registration phase, a training phase can take place. The user can try a certain set of authentication sessions, in order to familiarize with the system. Furthermore, the training phase is used to check if the system actually selects the expected kind of events.

In fact, consider our example in Sect. 4. There are plenty of possible meaning that could be associated to *a cat walking on the roof*: thinking to different subjects (different cat, different roof, different time of the day, etc.) or on the basis of different meanings, puns, metaphors or homonymies, for example: showing Cary Grant, acting as "the Cat" walking on the roof in a scene of the Alfred Hitchcock's "To Catch a Thief" (Hitchcock 1955). This effect is due to the fact that we assume the events database $\mathcal{S}$ is populated by a third party and that the event tagging process is accomplished in a way that may result unpredictable to the user (see Section 6).

Therefore, the training phase is required in order to tune the behavior of the system. To this end, the system could allow the user to refine the queries on her own or, alternatively, it could try to refine autonomously user's queries by asking her to accept/discard each event selected during the training session.

Periodically, the user can modify her secret by changing or adding pass-events, (*e.g.,* by modifying or adding queries in Q) by adding/replacing symbols in R or modifying mappings.

Authentication. During the authentication phase, the user sends her id u to the authenticator and waits for the challenge. The authenticator edits the challenge with n events in $\mathcal{S}$. To this end, it first retrieves k pass-events by running k randomly chosen queries from Q and second, it asks for $n - k$ random events which do not match any query in Q. The authenticator shows the challenge to the user through a video stream. The user composes the response by choosing for each pass-event e_i she recognize the corresponding response symbols $r_i = M'(e_i)$. Having received the response $r_1, \ldots, r_k$ by the user, the authenticator verifies whether $r_i = M(q_i), \forall (1 \leq i \leq k)$. In such a case, the authentication succeeds.

6 On the space of events

The space of events is a critical component in our architecture. Intuitively, the database is required to satisfy the following properties. Firstly, it is necessary that the database is big enough to ensure that the number of events that match any query is sufficiently high to avoid that certain videos appear too frequently in different challenges.

Secondly, all videos should be richly tagged according to homogeneous and stable criteria and users should be aware of these criteria. Otherwise, on the basis of any user's query, the system could select events that the user might not recognize. Indeed, we stress that FilmPW is based on the assumption that during the authentication of user u with secret $S(u) = (Q, R, M)$, for each pass event e_i in the challenge, it holds that $M(q_i) = M'(e_i)$ with a high probability. This is possible only if the user is able to recognize in the event e_i the meaning she "described" in q_i. Moreover, the tagging process should be reasonably stable, that is: tag assignments to a certain event should be almost stable/immutable over time. For example, we are incline to discourage/forbid tag removing since it may happen that due to tag removals, events originally considered as pass-events by the user may turn out to be distracting events for the system.

Finally, we argue that, on the long run, an adversary who collects a considerable amount of authentication sessions could successfully correlate the occurrence of certain events with the occurrence of certain responses, without being necessarily able to "understand" the meaning of events. This possibility is not so far from reality if one considers that nowadays, for example, free email providers as well as social networks perform millions of authentications per day through relatively insecure channels. To prevent such kind of attacks, the events stored in the database should have a sort of expiration time, therefore,

"expired" events should be constantly replaced with new ones.

These requirements make setup and operation of a really suitable EDBM a quite costly and resource-consuming task, especially in terms of manpower, due to the unavoidable human effort in the tagging process. Furthermore, we observe that such costs could grow with the number of users and with the time flowing, making the database management quite infeasible within a single authentication domain.

6.1 On self-producing and tagging events

Implementing and maintaining in-house the EDBM requires the Authentication Authority takes care of the management of the whole events life-cycle, i.e. creation/editing and tagging services as well as replacing expired videos. Events can be created by editing existing videos by means of semi-automatic editing tools. Video tools as `ffmpeg` (Bellard 2013), make possible to create sophisticated batch jobs that combine images, videos and audio files. Typically, events could be created starting from a sequence of pictures or existing events. Original videos could be created semi-automatically with tools used to create animated films. Well known software such as Blender (The Blender Foundation 2013) can be used to create small events in which characters and objects interact with each others. Such tools feature Python APIs that allow to create libraries of movements and interaction between 3D objects, though designing such objects and characters requires certain preliminary manual effort. The tagging process also requires a certain manual effort. Event tagging can leverage on several tools used to aid the video annotation in semantic video search services (Bertini et al. 2006). An increased automation degree can be reached by analyzing some implicit video characteristics (as in several video classification tools such as Mediamill (Worring et al. 2007) or by analyzing the audio track (e.g., with a speech recognition engine).

6.2 Leveraging on third party video-sharing services

We consider, as possible solution to overcome these problems, the use of existing Internet services for video sharing, such as YouTube, as EDBM, though many non trivial technical details should be further investigated. Nevertheless, it is possible to make some preliminary considerations according to the requirements we sketched above.

It seems rather difficult to obtain an authoritative assessment of the amount of videos stored by such file-sharing services. According the Wall Street Journal, in 2006, YouTube hosted about six million videos (Gomes 2006) whereas in 2013 this number has largely grown, considering that according to YouTube official statistics (YouTube LLC 2013): "More than 200 million videos have been claimed by Content ID". Anyway, we are evidently dealing with numbers that promise a sufficient variety of videos in terms of size, length, quality and topics.

According to the information published on the web site (YouTube LLC 2013), users upload to YouTube, about 72 h of video per minute. Videos are uploaded along with textual information (title, description) and an indefinite set of "tags", a sort of keywords that are used during search queries evaluation. Normally, these texts and tags are used to describe the video and give further information about its subjects and contents. Moreover, many videos also carry some comments, that may give further details. Normally, this information is entered using the natural language.

We still miss precise information about users' habits about tag management, but we are incline to consider it quite stable. Finally, although we are currently not able to estimate the length of videos life-cycle, it seems reasonable that the claimed upload rates should guarantee a considerable "freshness" of the event database.

However, we point out that a quantitative analysis of the data available, to give more concrete arguments in Cavour of our intuition, is one of the future direction of this research.

7 Preliminary usability tests

In this section we report the results of preliminary usability tests we have conducted with a small number of volunteers. Although these tests cannot be considered as a definite proof of usability, the system has been accepted by all volunteers very positively. The feeling we have had while running the tests is that users give higher attention to the challenge, w.r.t. "classical" (text-based) password schemes or even graphical passwords schemes (Catuogno and Galdi 2008). This impression has been confirmed by the users (through an interview whose results are summarized in Fig. 1) and has been justified by the inherent dynamism of the scheme. We first note that a user can identify a distracting event only at the end of the short film that represents it, since only at the end of the film she can claim that the short film does not represent any of her pass-events. On the other hand, she can identify a pass-event anytime during its representation. Since the challenge is a short video, all the users justified their attention as follows: if the current event is a distracting event, the user tries to identify it until the end of the current film frame; if the current event is a pass-event, once the user has identified it, she starts searching the end of the current frame and the beginning of the next one.

Fig. 1 The interview done to the users in order to rate the scheme acceptance

Statements	Do not agree	Partially agree	Agree	Score
Attention/effort required by authentication is greater than text-password schemes	0	4	16	56
This scheme looks suitable for authentication of every internet services	6	9	5	39
I would like to use this scheme for all my accounts	7	10	3	36
This scheme looks more secure than text-based password schemes	2	4	14	52
The authentication procedure is easy and comfortable	6	11	3	37
The authentication procedure is a pleasant experience	1	8	11	50
Authentication time is acceptable	6	12	2	36
The training phase is effective to make me familiar with the scheme	7	9	4	37
The training time is sufficient	2	11	7	45
Registration time is acceptable	13	6	1	28

The set of volunteers that participated in our usability test consisted of 20 persons with ages ranging from 20 to 53 years. Each user was asked to identify a secret consisting of 5 pass-events, to each of which she was asked to associate a string.

We developed a simple prototype composed by a set of scripts and a web application written in Python which features the registration phase, the training session and single authentications through a web-based interface. To setup the testbed, we have selected a dictionary of 500 words as $\mathcal{T}$, hence, we populated the database $\mathcal{S}$ by downloading about 200 videos from the Internet, chosen by running 100 queries composed by randomly chosen 5-tuples in $\mathcal{T}$ as tags. The set $\mathcal{R}$ contained strings of digits. Standard user secrets contained a set Q with seven queries, a set R with at least seven symbols and a mapping M with at least seven couples. Challenges were composed by five pass-events plus five distracting events.

The training phase was extremely simple. The users were required to execute three consecutive authentication sessions in the same day. The main problem experienced by the users during this phase was remembering the numbers associated to the events. A possible reason for this difficulty is that in our prototype we have a limited number of available concepts/actions. Thus, the users did not really create a *natural* association between the concept and the number but they were somehow forced to create such an association, with the results of randomly generated associations that are hard to remember. At the end of the training phase, 8 users out of 20 did not pass any of the three attempts they were allowed to use.

The users were asked to run (at most) three authentication sessions for one (resp., two) week after the training. In this case, the number of users that did not manage to authenticate decreased to 3, (resp., 1).

All the authentication sessions lasted 30–35 s, that is, essentially, the time needed to watch the whole challenge. A possible way to reduce the authentication time could be to compose the videos in the challenge in a matrix-like arrangement. We have asked the users to evaluate the system in terms of ease of use, access time and approval. All the users agreed that the experience had been extremely positive.

8 Conclusions

In this paper we have introduced FilmPW, a video-based authentication scheme. We have discussed some of the issues that arise when trying to create a taxonomy of human-recognizable actions/concepts. We have formally defined the authentication scheme and provided a prototype implementation. In order to validate our approach, we have conducted an initial test of usability. In our experience, the prototype has been well accepted by the users and its ease of use seems to lead to low error rates. A preliminary version of this work appeared in (Catuogno and Galdi 2013).

Acknowledgments The authors wish to thank Francesco Isgró for helpful discussions on image analysis and video event recognition.

References

Bellard F (2013) FFMPEG official web site. http://www.ffmpeg.org

Bertini M, Del Bimbo A, Torniai C, Cucchiara R, Grana C (2006) Mom: multimedia ontology manager. a framework for automatic annotation and semantic retrieval of video sequences. In: Proceedings of the 14th annual ACM international conference on Multimedia, ACM, pp 787–788

Bicakci K, Atalay N, Yuceel M, Gurbaslar H, Erdeniz B (2009) Towards Usable Solutions to Graphical Password Hotspot Problem. In: 2009 33rd Annual IEEE International Computer Software and Applications Conference, IEEE, pp 318–323

Blonder GE (1996) Graphical passwords. Lucent Technologies Inc, Murray Hill, NJ (US), US Patent no. 5559961

Blundo C, D'Arco P, Santis AD, Galdi C (2004) Hyppocrates: a new proactive password checker. J Syst Softw 71(1–2):163–175

Brezeale D, Cook DJ (2008) Automatic video classification: a survey of the literature. IEEE Trans Syst, Man, Cyber, Part C 38(3):416–430

Bursztein E, Martin M, Mitchell J (2011) Text-based captcha strengths and weaknesses. In: Proceedings of the 18th ACM conference on Computer and communications security, ACM, pp 125–138

Catuogno L, Galdi C (2008) A graphical pin authentication mechanism for smart cards and low-cost devices. In: Proceedings of the 2nd Workshop on Information Security Theory and Practices (WISTP 08) Sevilla (Spain), May 13–16, Springer-Verlag, Lecture Notes in Computer Science, vol 5019

Catuogno L, Galdi C (2010) On the security of a two-factor authentication scheme. In: Proceedings of the 4th Workshop on Information Security Theory and Practices (WISTP 2010) Passau (Germany), April 12–14, 2010, Springer, Lecture Notes in Computer Science, vol 6033

Catuogno L, Galdi C (2013) Towards the design of a film-based graphical password scheme. In: Information Science and Technology (ICIST), 2013 International Conference on, IEEE, pp 388–393

Catuogno L, Galdi C (2014) Analysis of a two-factor graphical password scheme. Intern J Inform Sec pp 1–17. doi:10.1007/s10207-014-0228-y

Ciaramella A, D'Arco P, De Santis A, Galdi C, Tagliaferri R (2006) Neural network techniques for proactive password checking. IEEE Trans Dependable Secure Compu 3(4):327–339

De Angeli A, Coventry L, Johnson G, Renaud K (2005) Is a picture really worth a thousand words? exploring the feasibility of graphical authentication systems. Intern J Human-comp Stud 63(1):128–152

De Luca A, Denzel M, Hussmann H (2009) Look into my eyes!: can you guess my password? In: Proceedings of the 5th Symposium on Usable Privacy and Security, ACM, p 7

Dhamija R, Perring A (2000) Dèjá vu: a user study using images for authentication. In: IX USENIX UNIX Security Symposium, Denver, Colorado (USA)

Gao H, Liu X (2009) A new graphical password scheme against spyware by using captcha. In: Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS 2009, Mountain View, California, USA, July 15–17, 2009, ACM, ACM International Conference Proceeding Series

Gibson M, Renaud K, Conrad M, Maple C (2009) Musipass: authenticating me softly with my song. In: Proceedings of the 2009 workshop on New security paradigms workshop, ACM, pp 85–100

Golle P (2008) Machine learning attacks against the asirra captcha. In: Proceedings of the 15th ACM conference on Computer and communications security, ACM, pp 535–542

Golle P, Wagner D (2007) Cryptanalysis of a cognitive authentication scheme (extended abstract). In: IEEE Symposium on Security and Privacy, IEEE Comp Soc, pp 66–70

Gomes L (2006) Will all of us get our 15 minutes on a youtube video? The Wall Street Journal online, August 30, 2006

Grady CL, Mcintosh AR, Rajah MN, Craik FIM (1998) Neural correlates of the episodic encoding of pictures and words. Proc Natl Acad Sci USA 95:2703–2708

Haller NM (1994) The S/KEY one-time password system. In: Proceedings of the Symposium on Network and Distributed System Security, pp 151–157

Harada A, Isarida T, Mizuno T, Nishigaki M (2006) A user authentication system using schema of visual memory. In: Biologically Inspired Approaches to Advanced Information Technology: Second International Workshop, Bioadit 2006, Osaka, Japan 26–27, 2006, Proceedings, Springer, Lecture Notes in Computer Science, vol 3853, pp 338–345

Hayashi E, Dhamija R, Christin N, Perrig A (2008) Use your illusion: Secure authentication usable anywhere. Proceedings of the 4th symposium on Usable privacy and security. ACM New York, NY, USA, pp 35–45

Hitchcock A (1955) To catch a thief. http://www.imdb.com/title/tt0048728/

Hopper NJ, Blum M (2001) Secure Human Identification Protocols. In: ASIACRYPT 2001, Springer, Lecture Notes in Computer Science, vol 2248, pp 52–66

Hoque E, Hoeber O, Strong G, Gong M (2013) Combining conceptual query expansion and visual search results exploration for web image retrieval. J Amb Intell Human Compu 4(3):389–400, http://www.scopus.com/inward/record.url?eid=2-s2.0-84878537451&partnerID=40&md5=a14779b5761ae42396369f31fec49759, cited By (since 1996)2

Jameel H, Shaikh R, Lee H, Lee S (2006) Human identification through image evaluation using secret predicates. Lect Notes Comp Sci 4377:67

Jensen W, Gavrila S, Korolev V, Ayers R, Swanstrom R (2003) Picture password: a visual login technique for mobile devices. In: National Institute of Standards and Technologies Interagency Report, vol NISTIR 7030

Jermyn I, Mayer A, Monrose F, Reiter MK, Rubin AD (1999) The design and analysis of graphical passwords. In: Proceedings of the 8th USENIX security Symposium, Washington

Jones MJ, Viola P (2001) Robust real-time object detection. In: Workshop on Statistical and Computational Theories of Vision, vol 266

Ko T (2008) A survey on behavior analysis in video surveillance for homeland security applications. In: AIPR, IEEE Comp Soc, pp 1–8

Kumar M, Garfinkel T, Boneh D, Winograd T (2007) Reducing shoulder-surfing by using gaze-based password entry. In: Symposium On Usable Privacy and Security (SOUPS)

Lanat A, Valenza G, Scilingo E (2013) Eye gaze patterns in emotional pictures. J Ambi Intell Human Compu 4(6):705–715

Lavee G, Rivlin E, Rudzsky M (2009) Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video.IEEE Trans Syst, Man, Cybern, Part C 39(5):489–504

Li S, Shah S, Khan M, Khayam S, Sadeghi A, Schmitz R (2010) Breaking e-banking CAPTCHAs. In: Proceedings of the 26th Annual Computer Security Applications Conference, ACM, pp 171–180

Maetz Y, Onno S, Heen O (2009) Recall-a-story, a story-telling graphical password system. In: Proceedings of the 5th Symposium on Usable Privacy and Security, ACM, p 27

Matsumoto T (1996) Human-computer cryptography: An attempt. In: ACM Conference on Computer and Communications Security, pp 68–75

McDonald DL, Atkinson RJ, Metz C (1995) One time passwords in everything (OPIE): Experiences with building and using stronger authentication. In: Fifth USENIX UNIX Security Symposium, Salt Lake City, Utah (USA)

Merler M, Huang B, Xie L, Hua G, Natsev A (2012) Semantic model vectors for complex video event recognition. IIEEE Trans Multimed 14(1):88–101

Real User Coorp (1998) Pass faces. http://www.realuser.com

Roth V, Richter K, Freidinger R (2004) A pin-entry method resilient against shoulder surfing. CCS '04: Proceedings of the 11th ACM conference on Computer and communications security. ACM Press, New York, NY, USA, pp 236–245

Ryoo MS, Chen CC, Aggarwal JK, Roy-Chowdhury A (2010) An overview of contest on semantic description of human activities (sdha) 2010. In: Proceedings of the 20th International conference on Recognizing patterns in signals, speech, images, and videos, Springer-Verlag, Berlin, Heidelberg, ICPR'10, pp 270–285, http://dl.acm.org/citation.cfm?id=1939170.1939208

Salehi-Abari A, Thorpe J, van Oorschot P (2008) On purely automated attacks and click-based graphical passwords. Proceedings of the 2008 Annual Computer Security Applications Conference. IEEE Computer Society, Washington, DC, USA, pp 111–120

Sasamoto H, Christin N, Hayashi E (2008) Undercover: authentication usable in front of prying eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp 183–192

Snoek CGM, Worring M (2005) Multimodal video indexing: A review of the state-of-the-art. Multimed Tools Appl 25(1):5–35. doi:10.1023/B:MTAP.0000046380.27575.a5

Sobrado L, Birget JC (2002) Graphical password. "The Rutgers Scholar, an electronic Bulletin for undergraduate research" 4

Suo X, Zhu Y, Owen GS (2005) Graphical passwords: a survey. In: Proceedings of 21st Annual Computer Security Application Conference (ACSAC 2005) december 5–9, Tucson AZ (US), pp 463–472

The Blender Foundation (2013) Blender official web site. http://www.blender.org

Thorpe J, van Oorschot P (2007) Human-seeded attacks and exploiting hot-spots in graphical passwords. In: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium table of contents, USENIX Association Berkeley, CA, USA

Tompkins DAD, Hoos HH (2004) UBCSAT: An implementation and experimentation environment for SLS algorithms for SAT and MAX-SAT. In: Proceedings of the Seventh International Conference on Theory and Applications of Satisfiability Testing (SAT 2004), pp 37–46

Weinshall D (2006) Cognitive authentication schemes safe against spyware (short paper). In: IEEE Symposium on Security and Privacy, IEEE Computer Society, pp 295–300

Wiedenbeck S, Waters J, Birget J, Brodskiy A, Memon N (2005) PassPoints: Design and longitudinal evaluation of a graphical password system. Intern J Human-Comp Stud 63(1–2):102–127

Wiedenbeck S, Waters J, Sobrado L, Birget JC (2006) Design and evaluation of a shoulder-surfing resistant graphical password scheme. In: Proceedings of Advanced Visual Interfaces AVI 2006, Venice ITALY

Worring M, Snoek CG, De Rooij O, Nguyen G, Smeulders A (2007) The mediamill semantic video search engine. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, IEEE, vol 4, pp IV-1213

Yan J, El Ahmad AS (2008) A low-cost attack on a microsoft captcha. In: Proceedings of the 15th ACM conference on Computer and communications security, ACM, pp 543–554

YouTube LLC (2013) Youtube fact sheet. http://www.youtube.com/t/fact_sheet